

**A Thesis Submitted for the Degree of PhD at the University of Warwick**

**Permanent WRAP URL:**

<http://wrap.warwick.ac.uk/95550/>

**Copyright and reuse:**

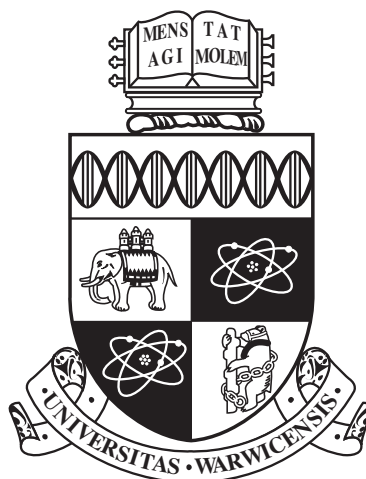
This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it.

Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: [wrap@warwick.ac.uk](mailto:wrap@warwick.ac.uk)



**Mechanistic Mathematical Models for the Design of  
Synthetic Biological Systems: DNA Recombination,  
Recombinase-Based Temporal Logic Gates and  
Antibiotic Production**

by

**Jack Edward Bowyer**

**Thesis**

Submitted to the University of Warwick

for the degree of

**Doctor of Philosophy**

**School of Engineering**

January 2018

THE UNIVERSITY OF  
**WARWICK**

# Contents

<b>Acknowledgments</b>	<b>iv</b>
<b>Declarations</b>	<b>v</b>
<b>Abstract</b>	<b>vi</b>
<b>Abbreviations</b>	<b>vii</b>
<b>Chapter 1 Introduction</b>	<b>1</b>
1.1 Main contributions of the thesis . . . . .	1
1.2 Publications arising from this research . . . . .	3
<b>Chapter 2 Mathematical Modelling in Synthetic Biology</b>	<b>4</b>
2.1 An introduction to synthetic biology . . . . .	4
2.1.1 The central dogma of molecular biology . . . . .	5
2.1.2 An overview of recent research in synthetic biology . . . . .	9
2.2 An introduction to mathematical modelling . . . . .	17
2.2.1 Mathematical models in synthetic biology . . . . .	18
2.2.2 The trade-off between mechanistic and black box modelling approaches . . . . .	20
2.3 Mathematical modelling techniques . . . . .	23
2.3.1 Ordinary differential equations and initial value problems . .	23
2.3.2 The law of mass action . . . . .	27
2.3.3 Explicit and numerical solutions to initial value problems . .	28
2.3.4 Model reduction . . . . .	34
2.3.5 Non-dimensionalisation . . . . .	37
2.4 Parameter inference . . . . .	40
2.4.1 Global optimisation . . . . .	40
2.4.2 Approximate Bayesian computation . . . . .	45

<b>Chapter 3 Mechanistic Modelling of a Rewritable Recombinase Addressable Data Module</b>	<b>54</b>
3.1 Scientific background . . . . .	54
3.1.1 DNA recombination . . . . .	54
3.1.2 Existing recombinase-based systems . . . . .	58
3.1.3 Existing recombinase-based models . . . . .	60
3.2 Formulating a mechanistic model of <i>in vitro</i> RAD module dynamics	64
3.3 Model validation via global optimisation . . . . .	72
3.3.1 Experimental methods . . . . .	78
3.4 Non-dimensional simulations of <i>in vivo</i> RAD module dynamics . . .	78
3.5 Conclusions . . . . .	87
 <b>Chapter 4 Mechanistic Modelling of a Recombinase-Based Two-Input Temporal Logic Gate</b>	 <b>89</b>
4.1 Scientific background . . . . .	89
4.1.1 Boolean algebra and logic gates . . . . .	89
4.1.2 Existing logic gate systems . . . . .	94
4.1.3 Existing logic gate models . . . . .	98
4.2 Formulating an <i>in vivo</i> reaction network of a recombinase-based temporal logic gate . . . . .	100
4.3 Constructing a mechanistic model of the temporal logic gate . . . . .	101
4.4 Model validation via global optimisation . . . . .	106
4.4.1 Experimental methods . . . . .	111
4.5 Reversing the roles of integrase inputs . . . . .	113
4.6 Conclusions . . . . .	115
 <b>Chapter 5 DNA Recombination Experiments</b>	 <b>117</b>
5.1 Establishing ideal bacterial growth conditions . . . . .	117
5.2 Recording excision time courses . . . . .	123
5.3 Conclusions . . . . .	129
5.4 Methods . . . . .	130
5.4.1 Intergenic conjugation protocol . . . . .	130
5.4.2 Media stock solutions . . . . .	131
 <b>Chapter 6 Mechanistic Modelling of the Regulatory System Controlling Methylenomycin Production in <i>Streptomyces coelicolor</i></b>	 <b>132</b>
6.1 Scientific background . . . . .	132
6.1.1 A brief history of antibiotics . . . . .	132

6.1.2	<i>Streptomyces</i> . . . . .	135
6.1.3	The methylenomycin regulatory gene cluster . . . . .	136
6.2	Formulation of candidate model architectures . . . . .	138
6.3	Available experimental data . . . . .	142
6.4	Model selection via approximate Bayesian computation . . . . .	144
6.5	Parameter inference via global optimisation . . . . .	147
6.6	Monte Carlo simulations of methylenomycin production in mutant strains . . . . .	149
6.7	Experimental design for future studies . . . . .	152
6.8	Conclusions . . . . .	153
<b>Chapter 7 Conclusions and Future Work</b>		<b>154</b>
7.1	Conclusions and discussion . . . . .	154
7.2	Future work . . . . .	159

# Acknowledgments

I would like to firmly acknowledge my supervisor Declan Bates whose guidance throughout my PhD course has been completely invaluable and to whom I will always be grateful for giving me the opportunity to do the research I wanted on my own terms.

I owe many thanks to a number of people who have helped me, and have made my PhD a thoroughly enjoyable experience. Thanks to Rucha Sawlekar and Mel du Lac for being fantastic friends and colleagues; to Anup Das for your wise words and for being so accommodating; to Kathryn Styles for teaching me everything I know about biology and for having the patience of a saint, and to Christophe Corre for letting me loose in your lab and for being generally brilliant. My thanks go to everyone else that I have collaborated with over the last few years, this thesis would not be the same without your input and thank you, in particular, to Wilson Wong for helping to establish my next research opportunity.

I also want to thank Mum, Dad, Bobby and the rest of my family for always being so incredibly supportive and continuing to ask me about my research whilst nodding and smiling so genuinely.

Most of all, I want to thank my lovely Becky without whom I would never have found the perfect PhD and the perfect base to work from. Your bravery in the face of living with someone on a PhD salary has been nothing short of phenomenal, thank you.

Thank you EPSRC for said salary.

# Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for a degree.

The research presented was carried out by the author with the exception of the cases outlined below:

- The experimental data described in Chapter 3 were generated by Jia Zhao, University of Glasgow.
- The experimental data described in Chapter 4 were generated by Victoria Hsiao, California Institute of Technology.
- The experimental protocols followed in Chapter 5 were provided by Kathryn Styles, University of Warwick.

The research presented in Chapters 3, 4 and 6 of this thesis has been published and/or submitted for publication by the author. Full details of all five papers published or submitted as a result of this research are given in Chapter 1.

# Abstract

Synthetic biology is the design and implementation of novel biological devices via the application of engineering principles to biological systems research. Mathematical modelling is an invaluable tool in developing our understanding of biological system dynamics and characterising small parts and circuits for the assembly of higher-level systems.

In this thesis, mathematical modelling approaches are applied to three biological circuits of interest. A novel mechanistic model of the DNA recombination reactions comprising a genetic switch reveals the input criteria and operational specifications required of a digital data storage module. Specific layering of the components comprising recombinase-based genetic switches can provide cellular Boolean logic operations. A novel mechanistic model of a two-input temporal logic gate is able to simulate and predict *in vivo* dynamical responses captured by a large experimental dataset. Experimental implementation of recombinase-based circuitry is unpredictable and can lead to lengthy development times, providing clear evidence of the advantages of utilising mathematical models in synthetic biology.

Antibiotic resistance has become one of the most prominent challenges facing medicine today, placing immense importance on the characterisation of new natural products. The first detailed mathematical model of the methylenomycin A producing gene cluster in the bacterium *Streptomyces coelicolor* is developed through the application of model selection to a large set of candidate system architectures.

Mathematical models presented in this thesis can be adapted and expanded to suit many different experimental conditions and system responses, facilitating the design of novel synthetic biological circuitry.



# Abbreviations

ABC – approximate Bayesian computation.  
ABM – agent-based modelling.  
BDF – backward differentiation formula.  
BLADE – Boolean logic and arithmetic through DNA excision.  
BSA – bovine serum albumin.  
Caltech – California Institute of Technology.  
CCD – charge coupled device.  
CHO – Chinese hamster ovary.  
DIC – DNA invertase cascade.  
DNA – deoxyribonucleic acid.  
DSRS – dynamic sensor-regulator system.  
*E. coli* – *Escherichia coli*.  
EDTA – ethylenediamin etetraacetic acid.  
FLP – flippase.  
GA – genetic algorithm.  
GFP – green fluorescent protein.  
HGT – horizontal gene transfer.  
H1MF – H1 multi-format.  
iGEM – international genetically engineered machine.  
IL-12 – interleukin-12.  
IQR – interquartile range.  
IVP – initial value problem.  
MARE – methylenomycin regulatory element.  
MDR – multidrug resistant.  
MIT – Massachusetts Institute of Technology.  
MMF – methylenomycin furan.  
MMY – methylenomycin.  
mRNA – messenger RNA.  
MRSA – methicillin-resistant staphylococcus aureus.  
NAND – not AND.  
NDF – numerical differentiation formula.  
N-IMPLY – not IMPLY.  
NOR – not OR.  
NOTIF – not IF.

OD – optical density.  
ODE – ordinary differential equation.  
PDE – partial differential equation.  
RAD – recombinase addressable data.  
RDF – recombination directionality factor.  
RFP – red fluorescent protein.  
RKDP – Runge-Kutta Dormand-Prince.  
RK2 – second order Runge-Kutta.  
RK4 – fourth order Runge-Kutta.  
RNA – ribonucleic acid.  
RNAi – RNA interference.  
rpm – revolutions per minute.  
rRNA – ribosomal RNA.  
RSBP – Registry of Standard Biological Parts.  
RTC – riboregulated transcriptional cascade.  
SBML – Synthetic Biology Markup Language.  
SB1.0 – the first international meeting on synthetic biology.  
*S. cerevisiae* – *Saccharomyces cerevisiae*.  
*S. coelicolor* – *Streptomyces coelicolor*.  
SDE – stochastic differential equation.  
SFM – soya-flour mannitol.  
SMC – sequential Monte Carlo.  
SSR – site-specific recombinase.  
TDR – totally drug resistant.  
TMTC – too many to count.  
tRNA – transfer RNA.  
UV – ultraviolet.  
XDR – extremely drug resistant.  
XML – Extensible Markup Language.  
XOR – exclusive OR.

# Chapter 1

## Introduction

### 1.1 Main contributions of the thesis

This thesis presents three mathematical modelling investigations relating to biological systems that represent potentially fundamental components in the assembly of novel synthetic circuits. Characterisation of such circuits is often an *ad hoc* and time consuming process if performed through experimentation alone. Hence, synthetic biology approaches are increasingly reliant on predictive mathematical models in developing understanding of desirable dynamical responses. The modelling investigations presented in this thesis are focused on DNA recombination and the regulation of antibiotic production and are mechanistic in nature; considerable efforts are made to account for maximal biological detail wherever possible. The content of the thesis is delivered with the following general structure:

Chapter 2 consists of an introduction to the field of synthetic biology and the relevant mathematical modelling techniques that facilitate the analysis of the biological systems investigated in this thesis. The motivation behind the application of mathematical modelling approaches in synthetic biology is discussed, highlighting the importance of interdisciplinary communication and collaboration in the development of the field.

Chapter 3 consists of the research relating to the characterisation of the recombinase-based genetic toggle switch which we have come to refer to as a rewritable recombinase addressable data (RAD) module. Although site-specific recombinases (SSRs) are attracting increasing attention as reliable DNA manipulation tools, there is not currently a wealth of mathematical modelling investigations concerning the operational properties of the associated circuitry. In this chapter, we identify optimal switching profiles through the construction and experimental validation of a

mechanistic model of the RAD module.

Chapter 4 presents an extension of the research presented in Chapter 3 through a mechanistic modelling investigation relating to a recombinase-based two-input temporal logic gate. Standard logic functions have been identified through several transcriptional and recombinase-based circuits however, temporal logic gates are not well documented in the literature. Our mechanistic model is comprised of multiple integration pathways taken from our validated RAD module model and provides quantitative simulations of the available experimental data.

Chapter 5 describes the experimentation carried out to develop understanding of DNA recombination and experimental data collection. The objective of these experiments is to examine the efficiency of the integrase-excisionase-mediated excision (deletion) of a gene coding for bioluminescence in the bacterium *Streptomyces coelicolor*. This work highlights a number of limitations with current experimental protocols, which are currently still at an early stage of maturity. However, it also provides some valuable insights into the underlying biology of recombinases, and suggests ways in which these systems could be further developed in future studies.

Chapter 6 consists of the research relating to the identification of the most plausible modelling architecture of the regulatory system controlling methylenomycin production in the model organism *Streptomyces coelicolor*. Although homologous regulatory systems have been identified in numerous microorganisms, no attempt has been made to characterise methylenomycin regulation *in silico*. Our model architecture is identified through a probabilistic model selection approach and is optimised against experimental time course data. Model simulations can also replicate qualitative experimental observations regarding mutant bacterial strains.

Chapter 7 describes the overall conclusions arising from the research conducted in this thesis and discusses the potential of future work in expanding the impact of novel synthetic systems comprising the circuits investigated. We summarise the results presented in each chapter and discuss some of the broader implications these have on future research. We discuss the benefits of extending model reduction efforts in order to facilitate extensive mathematical analysis of our mechanistic models in the future. A number of opportunities for further research are presented, such as developing our temporal logic gate model in order to examine the functional differences between distinct integrases, and the effect that altering the roles of these integrases has on the output of the logic gate. We also outline plans for further collaborative work relating to mechanistic modelling and experimental implementation of higher-level systems in mammalian cell lines.

Synthetic biology is reliant upon the shared expertise of biologists and engi-

neers in collaborative research efforts. Effective communication of information and ideas is fundamental to bridging the gap between these disciplines. To this end, the introductory material in this thesis is written in a style that is intended to be accessible to biologists encountering advanced mathematics, and to mathematicians and engineers encountering advanced cellular biology.

## 1.2 Publications arising from this research

- J. Bowyer, J. Zhao, S. Rosser, S. Colloms and D. Bates. Development and experimental validation of a mechanistic model of *in vitro* DNA recombination. In *Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Milano, Italy, August 2015, pages 945-948, doi: 10.1109/EMBC.2015.7318519. (relating to Chapter 3)
- J. Bowyer, J. Zhao, P. Subsoontorn, W. Wong, S. Rosser and D. Bates. Mechanistic modelling of a rewritable recombinase addressable data module. *IEEE Transactions on Biomedical Circuits and Systems*, 10(6):1161-1170, December 2016, doi: 10.1109/TBCAS.2016.2526668. (relating to Chapter 3)
- J. Bowyer, V. Hsiao and D. Bates. Development and experimental validation of a mechanistic model of a recombinase-based temporal logic gate. In *Proceedings of the 12th IEEE BioMedical Circuits and Systems Conference*, Shanghai, China, October 2016, pages 464-467, doi: 10.1109/BioCAS.2016.7833832. (relating to Chapter 4)
- J. Bowyer, V. Hsiao, W. Wong and D. Bates. Mechanistic modelling of a recombinase-based two-input temporal logic gate. *IET Engineering Biology*, 1(1):40-50, July 2017, doi: 10.1049/enb.2017.0006. (relating to Chapter 4)
- J. Bowyer, E. de los Santos, K. Styles, A. Fullwood, C. Corre and D. Bates. Modeling the architecture of the regulatory system controlling methylenomycin production in *Streptomyces coelicolor*. *BMC Journal of Biological Engineering*, July 2017. (relating to Chapter 6)

## Chapter 2

# Mathematical Modelling in Synthetic Biology

### 2.1 An introduction to synthetic biology

Synthetic biology is a highly interdisciplinary field of research with a primary focus on the application of engineering principles to biological systems. The characterisation of small biological parts and circuits is essential in providing the building blocks with which to assemble novel devices. In this context, biological parts are the genetic elements that mediate the natural processes of life. Rigorous analysis of natural systems and mechanisms that have evolutionary prominence will facilitate the design and construction of biological modules capable of potentially unlimited user-defined outputs. From an analytical perspective, synthetic biology is fundamentally comprised of mathematical biology approaches to examining the dynamical behaviour of biological systems through a mathematical framework. This premise undergoes a natural progression towards systems biology research whereby analysis of specific mechanistic properties is able to reveal performance criteria and other functional insights. The motivation behind synthetic biology, and the reason why the field exists in its own right, is the idea that standardised biological parts can be assembled to form small circuitry and larger-scale modules in a manner analogous to electrical circuits and machinery. Classical engineering principles such as system design and optimisation contribute significantly to this research effort, however, as the name implies, synthetic biology is, first and foremost, a biology discipline. Hence, the natural flow of genetic information exists at the core of engineering biological systems.

### 2.1.1 The central dogma of molecular biology

In order to highlight the relevance of the central dogma of molecular biology within synthetic biology, it is necessary to review the biology of the cell. There are many different types of cell, appearing across all domains and kingdoms of organisms, that consist of many different types of organelle [Nicholl, 2008]. An organelle is a cellular component that provides a specific function, akin to organs in the human body. Despite this remarkable variety in their internal composition, all cells are united, at least initially, by the presence of genetic material (usually DNA) that provides the information necessary for the cell to function. The broad classification of cells relates to the internal disposition of the DNA; in eukaryotic cells, most notably animal and plant cells, DNA is contained within a membrane-bound nucleus whereas, in prokaryotic cells such as bacteria, the DNA is not membrane-bound (Figure 2.1). Additional internal structure is provided by other membranes in eukaryotic cells,

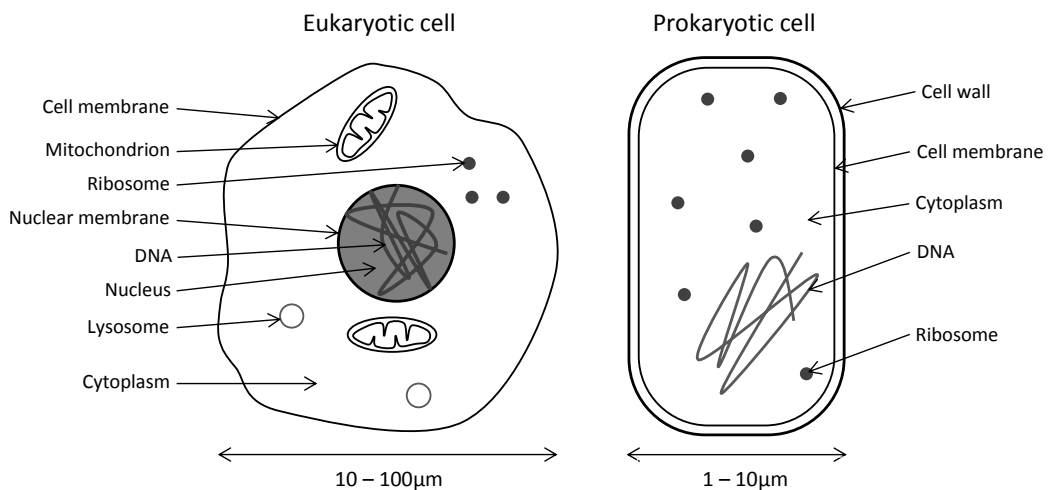


Figure 2.1: Schematic diagram of a eukaryotic cell and a prokaryotic cell. Cell membranes, ribosomes and DNA are common to both cells however, these two cellular compositions are largely distinct in general; some detail is omitted here for the sake of brevity.

which are also typically larger than prokaryotic cells [Alberts et al., 2002]. There exist upper and lower limits on the optimal size of cells [Marshall et al., 2012] since essential gas and nutrient exchange is mediated by diffusion, which is dependent on the dimensions of the molecules involved in this process. It is an increase in the total number of cells that permits the increase in the size of multicellular eukaryotes, not an increase in individual cell size. Mitochondria, ribosomes and lysosomes are the organelles responsible for cellular respiration, protein synthesis and waste removal

respectively, making them vital to the maintenance of the cell.

The importance of DNA cannot be underestimated given the intrinsic relationship it shares with cellular classification and functionality. That is, the functional properties of a cell are defined by its DNA, which provides the instructions for the synthesis of proteins which, in turn, manage the processes of life. Put simply, DNA is the blueprint of life and there are three main criteria that allow it to fulfil such a crucial biological role [Nicholl, 2008]. Firstly, DNA must be able to produce identical copies of itself through replication in order for genetic material to be passed on during the growth and development of new cells. Secondly, the stability of the molecule must be sufficiently high to maintain the regular functionality of genetic material for many years in living organisms. Thirdly, the effects of evolutionary pressures must be able to induce small alterations in the genetic code, referred to as mutations, to allow organisms to survive and adapt to changes in their immediate environments. These criteria are satisfied largely by the molecular composition of DNA which reveals the ingenuity of complementary base pairing and, consequently, the language of genetics.

Deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), as their names suggest, are nucleic acids. They are comprised of monomeric components known as nucleotides and chains of these nucleotides, called polynucleotides, give nucleic acids their overall structure. Three smaller components comprise the nucleotides themselves: a sugar, a phosphate group and a nitrogenous base. It is the sugar component that defines the nucleotide, 2'-deoxyribose in DNA and ribose in RNA, and this also determines the structural characteristics of the polynucleotide along with the phosphate group. DNA adopts a double stranded, helical structure that was first proposed by James Watson and Francis Crick in 1953 [Alberts et al., 2002]. The characteristics of the polynucleotide that relate to genetic information dissemination and storage are provided by the nitrogenous bases which therefore have great importance in relation to the coding function of nucleic acids. There are four distinct nitrogenous bases in DNA, namely, adenine (A), guanine (G), cytosine (C) and thymine (T); the thymine base is replaced by the functionally equivalent uracil (U) in RNA. These bases form the language of genetics by virtue of a 'dictionary' of combinations that code for specific amino acids. Adenine and guanine are classified as purines, which have a double-ring chemical structure, whereas cytosine, thymine and uracil are classified as pyrimidines, which have a single-ring chemical structure. Base pairing is key to encoding and decoding genetic information and is strictly conditional on one particular arrangement, that is, A paired with T (or U) and G paired with C. The only satisfactory arrangement is a purine-pyrimidine base pair



due to the structural limitations of the nucleic acid and the atomic composition of the bases themselves. An A-T base pair is secured by two hydrogen bonds whereas a G-C base pair is secured by three hydrogen bonds. The four letter alphabet of bases (A, G, C and T) defines a language of three-letter ‘words’ that code for the requisite molecules that drive gene expression. Genetic information is ultimately expressed via proteins, a significant subset of which constitute the enzymes that catalyse metabolic reactions.

All natural proteins are comprised of unique sequences of amino acids called polypeptides, synthesised from a set of twenty distinct amino acids [Nicholl, 2008]. The genetic sequences that code for individual amino acids, known as codons, are three bases in length. The reason for this is that if the bases were used singly to code for individual amino acids, there would only be four possible choices and hence a maximum of only four amino acids that could be synthesised. There still wouldn’t be enough permutations to code for twenty amino acids if the bases were coded in pairs, since only sixteen distinct choices are available ( $4^2 = 16 < 20$ ). However, triplets of bases are able to provide the sufficient permutations ( $4^3 = 64 > 20$ ) and these triplets form the three-letter words that code for each amino acid. Hence, the minimum coding requirement on a DNA strand for a protein consisting of 100 amino acids would be 300 nucleotides. Utilising codons that are three bases in length clearly provides more words in the dictionary than are actually necessary; there are three codons that signal the termination of protein synthesis, referred to as stop codons, and the remaining codons are accounted for by the fact that groups of distinct codons are able to code for the same amino acid, referred to as a redundancy in the code.

The chemical structure of RNA differs from that of DNA in the sense that it is most commonly a single stranded molecule. There are three main classes of RNA molecules, the most abundant of which is ribosomal RNA (rRNA), making up approximately 85% of total cellular RNA. It is the RNA component of ribosomes, which is essential in translating the information necessary for gene expression. Transfer RNA (tRNA) provides amino acids with the specificity required for correct insertion into the polypeptide undergoing synthesis and comprises approximately 10% of total RNA. Genetic information is transported through the nuclear membrane to the ribosome by messenger RNA (mRNA) which usually comprises less than 5% of total cellular RNA [Voit, 2013].

Given the immense significance of DNA and genetic dissemination across the field of biology, the central dogma of molecular biology (Figure 2.2) intuitively constitutes a basic overview of gene expression as follows: gene expression is initiated

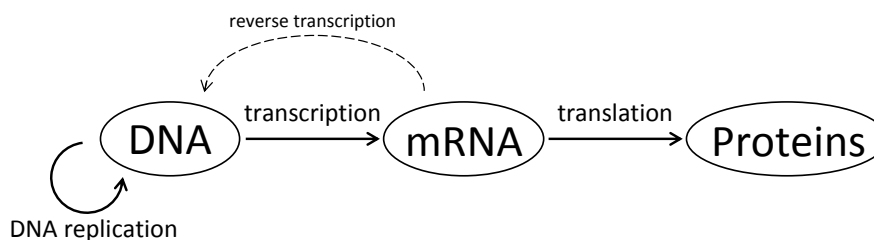


Figure 2.2: Schematic diagram representing the central dogma of molecular biology. Genetic material is disseminated in a unidirectional manner with the exception of reverse transcription. Figure adapted from Nicholl [2008].

by transcription, the process in which an enzyme known as RNA polymerase converts DNA into a complementary mRNA strand. The mRNA is then transported through the nuclear membrane into the cytoplasm in order to carry out the second stage of gene expression, translation. When the mRNA encounters a ribosome, the ribosome translates the mRNA to form polypeptides which fold to produce active proteins capable of performing cellular functions. The passage of genetic information is unidirectional with the exception of reverse transcription whereby DNA is synthesised from RNA, for example, in the event of infection of a host cell by a retrovirus (RNA-based virus) [Nicholl, 2008].

Engineering biological systems inevitably constitutes a systematic analysis of the processes outlined in the central dogma. In order to realise specific engineered biological outputs it is necessary to alter this natural flow of cellular processes which, in itself, encapsulates much of what might be regarded as the essence of synthetic biology. A key concept that has been adopted widely across the field is that such processes can be regarded as input-output systems comprised of a wealth of biological machinery and components. It is important, therefore, that synthetic biologists are able to design novel systems using standardised components (parts) that perform as expected in the relevant biological contexts. The definition of standardisation in a biological context is analogous to that in modern engineering and manufacturing, whereby a universally recognised catalogue of components facilitates efficient module assembly and reliable functionality. There are a number of transcriptional elements that have already been standardised sufficiently for engineering synthetic circuitry [Purnick and Weiss, 2009]. These include promoters, sequences that are bound by RNA polymerase to initiate transcription; repressors, proteins that prevent transcription by binding to promoters instead of RNA polymerase; terminators, sequences that cause transcription to end; and ribosome binding sites (RBSs), mRNA sequences that recruit ribosomes to initiate translation. Standardisation of

biological parts is achieved through extensive experimentation and, in turn, well documented functional profiles that can be disseminated and applied universally. Standardised parts, known as BioBricks, are essential to the construction of novel synthetic systems with respect to the compatibility required in the design process and the ease of their physical assembly in model organisms. In this way, innovative, user-defined functionality can be realised through circuit design analogous to that of electronic circuits. The majority of the implementation of synthetic biological systems is currently performed in the model bacterium *Escherichia coli*. Over many years of research, *E. coli* has been fully characterised, and was one of the first microorganisms to have its genome sequenced, making it ideally suited to laboratory procedures and experimentation [Alberts et al., 2002]. That said, a host of other bacteria, yeasts and mammalian cells are also used to implement novel circuits based on their metabolic and growth characteristics, among others. With respect to human benefits such as disease therapeutics, and other biomedical solutions arising from the application of synthetic biological devices, implementation within eukaryotic cells will naturally be required in order to realise the potential efficacy of such systems.

### 2.1.2 An overview of recent research in synthetic biology

Synthetic biology is a relatively new field of research. The emergence of the fundamental concepts that have come to define the subject were first demonstrated at the turn of the millennium through two innovative synthetic systems, the ‘repressilator’ [Elowitz and Leibler, 2000] and the ‘toggle-switch’ [Gardner et al., 2000]. These two pioneering circuits provide clear evidence that novel functionality can be achieved through the assembly of biological parts, capturing the essence of synthetic biology and delivering the proof of concept that has unlocked the potential of the field. Furthermore, these systems were both published in conjunction with mathematical modelling investigations, illustrating the fundamental necessity within synthetic biology to exploit interdisciplinary research efforts and the predictive capabilities of mathematical models through computational simulation.

The repressilator is designed to produce regular oscillatory behaviour using components that are not found in any natural circadian clock, that is, a ‘body clock’ that operates on an oscillatory 24-hour period in tandem with the day-night cycle. Three repressor proteins, LacI, TetR and CI, comprise a cyclic negative feedback loop that exhibits temporal oscillations in the concentration of each of these protein components (Figure 2.3A). LacI inhibits the production of TetR which, in turn, inhibits the production of CI which then completes the cycle by inhibiting LacI

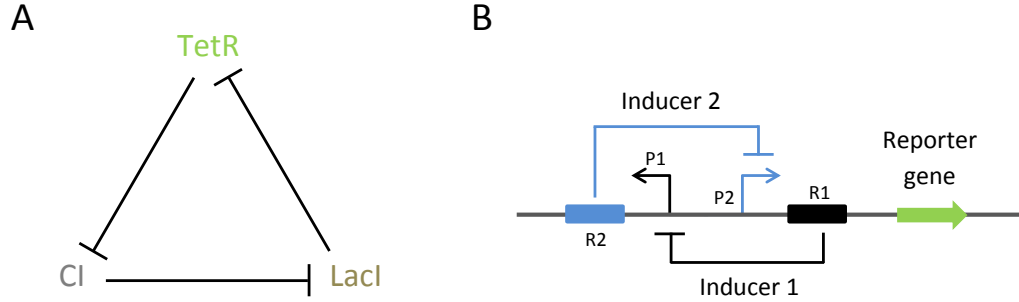


Figure 2.3: Schematic diagram representing the first synthetic biological circuitry. A) The repressilator gives rise to oscillatory behaviour through a mutually inhibitory arrangement of three repressor proteins. B) The genetic toggle switch gives rise to bistability through a mutually inhibitory arrangement of two repressible promoters. Figure adapted from Elowitz and Leibler [2000] and Gardner et al. [2000].

production. The TetR repressor is tagged with green fluorescent protein (GFP), with the levels of fluorescence being representative of TetR concentration in the cell, and thus provides observable periodicity as a readout of the system state over time. The two genetic constructs comprising the repressilator and the reporter are built into separate plasmids and transformed into *E. coli*.

A simple mathematical model of transcriptional regulation was used to design the repressilator. The model consists of six coupled equations representing the rate of change of the three repressor protein concentrations as well as their corresponding mRNA concentrations. Model simulations demonstrate that the system exhibits temporal oscillations in the concentrations of its three components, in alignment with the observed experimental dynamics. Model analysis reveals a unique steady state to which there are two possible solutions, depending on the reactions rates comprising the model parameter set. Firstly, the solution converges towards a stable steady state or, secondly, the steady state may become unstable leading to the desired sustained oscillations. The behaviour of dynamical systems is greatly influenced by the rates of the associated reactions, in this case transcription and translation rates as well as the rate of protein and mRNA degradation. The model assisted in identifying that oscillations are instigated by reduced transcriptional leakiness and similarity in protein and mRNA degradation rates. As a result, two alterations were made to increase the chance of the synthetic network functioning in the oscillatory regime. Firstly, hybrid promoters were used to address transcriptional efficiency and, secondly, a protein destruction marker was used to establish comparable effective lifetimes of the repressor proteins and mRNA.

Hence, the repressilator provides a near perfect example of the benefit of modelling insight on the direction of experimentation and the *in vivo* implementation of artificial networks that might otherwise have required far greater time frames to achieve through experimental trials. That said, the repressilator is also shown to exhibit noisier and more variable behaviour compared to that of natural circadian clocks and thus highlights a major challenge in synthetic biology concerning reliable and robust deployment of novel systems *in vivo* [Elowitz and Leibler, 2000].

The design of the genetic toggle switch is centred on bistability, whereby two inducible stable states form the basis of a synthetic gene-regulatory network. A mutually inhibitory arrangement of repressible promoters gives rise to the bistability that characterises the system, effectively providing an ‘on’ or ‘off’ switch-like response. Theoretically, this design could facilitate the expression of any gene of interest. The toggle switch is composed of two constitutive promoters (P1 and P2) each of which is positioned upstream of two repressors (R1 and R2) (Figure 2.3B). In the absence of induction, each promoter is able to transcribe the repressor corresponding to the opposite promoter however, transcription of the two repressors will cause inhibition of the corresponding promoters and nullify the output of the system. Induction of R1 will cause repression of P1 and will therefore nullify transcription of R2 which allows P2 to perform unrestricted transcription, thus establishing one of the two stable gene expression states. The second state is realised in a similar manner through induction of R2. A reporter gene *gfpmut3* is positioned downstream of P2, tagging this state with GFP in order to clearly monitor the readout of the system as a function of fluorescence intensity. Induction of R1 therefore results in the expression of GFP, termed the high state, and induction of R2 which inhibits GFP expression, i.e. no output, is termed the low state. Transient repressor induction is achieved chemically or thermally, giving the user temporal control of the output of the toggle switch without the need for continuous induction. As with the repressilator, the genetic constructs of the toggle switch are plasmid-based and transformed into *E. coli*.

The simple mathematical model of the toggle switch consists of two equations describing the rate of change in concentration of the two repressors and is parameterised in accordance with their cooperativity and rates of synthesis, two properties believed to be fundamental to the network. Cooperativity refers to the effect of ligand binding on the affinity of subsequent ligand binding at other sites on the same molecule, be it positive or negative. The model reveals the conditions necessary for bistability: balanced promoter strengths give rise to the one unstable and two stable steady states that provide the two basins of attraction required for the

desired bistability, whereas unbalanced promoter strengths produce just one stable steady state. The size of the bistable region is directly proportional to the magnitude of the cooperativity of repression. The toggle switch is designed to incorporate the fewest genes and regulatory elements possible in facilitating robust bistable behaviour. That is, the inherent variability of gene expression has a negligible effect on the two states, thus minimising the chance of random switching events, and the criteria for bistability comprise a relatively large portion of the parameter space, not limited to a small subset [Gardner et al., 2000].

To reiterate, as with the repressilator, the approach to engineering a genetic toggle switch is driven primarily by the manipulation of the network architecture which represents a substantial deviation from conventional genetic engineering practices that typically achieve desired outputs via the restructuring of the relevant regulatory elements themselves. The toggle switch again captures the essence of synthetic biology by demonstrating qualitative agreement between mathematical modelling and experimental observations. As an inducible gene regulatory circuit, the toggle switch forms a cellular memory unit which has an array of potential applications in biotechnology, biocomputing and gene therapy.

The repressilator and the toggle switch initiated what has become known as the first wave of synthetic biology [Purnick and Weiss, 2009]. This initial phase focused primarily on parts characterisation in order to accumulate a wide-ranging selection of BioBricks with which to engineer such novel systems. As a result, a number of small circuits were published, some being oscillators [Atkinson et al., 2003; Goh et al., 2008; Stricker et al., 2008; Swinburne et al., 2008; Tiggles et al., 2009] and switches [Kramer et al., 2004b; Atkinson et al., 2003; Ham et al., 2008] akin to the aforementioned pioneering studies as well as other innovative designs including pulse generators [Basu et al., 2005] and logic circuits [Win and Smolke, 2008; Rinaudo et al., 2007]. The first signs of extending research beyond simple gene regulatory networks are identified even during this early period with the advent of cell-cell communication circuits [You et al., 2004] and post-translational regulation [Park et al., 2003]. Difficulties arose regarding *ad hoc* approaches to circuit assembly and optimisation which resulted in time consuming, laborious iterations of experimentation to establish desired functionality. Hence, it became apparent that further progress would be dependent on large-scale standardisation, reiterating the importance of extensive parts characterisation.

By 2003, the international genetically engineered machine (iGEM) foundation was established [Cameron et al., 2014]. This independent, non-profit organisation promotes the advancement of synthetic biology and the development of a collab-

orative, interdisciplinary community. One of the main programs of the foundation is the annual iGEM competition, the first of which was held at the Massachusetts Institute of Technology (MIT) in 2004 and hosted five teams from American universities. Today, approximately 300 teams from more than 30 countries compete. The competition gives students of multiple disciplines the opportunity to design novel synthetic circuits using the catalogue of BioBricks currently available. These parts comprise one of the foundation’s other main programs, the Registry of Standard Biological Parts (RSBP). The registry currently consists of over 20,000 compatible BioBricks that provide an open source of parts during competition and for synthetic biology research across the community. New synthetic circuitry and BioBricks established by iGEM teams and other labs are added to the registry in order to expand the scope of the registry and field alike. The first international meeting on synthetic biology (SB1.0) was also held at MIT in 2003. This provided the platform for biologists, chemists, physicists, engineers, computer scientists and mathematicians to lay the foundations of a collective vision [Cameron et al., 2014].

As synthetic biology continued to gain momentum, a host of other breakthrough systems emerged [Cameron et al., 2014]. Quorum sensing, the population density-dependent response of a system, was engineered in *E. coli* in order to facilitate multicellular patterning [Basu et al., 2005; You et al., 2004]. Post-transcriptional and post-translational regulation was developed through RNA-based circuitry [Isaacs et al., 2003; Bayer and Smolke, 2005]. Specific Boolean logic operations such as AND gates were realised *in vivo* via engineering the tRNA required for translation [Anderson et al., 2006] and also in other studies through orthogonal ribosomes [Rackham and Chin, 2005]. The most notable breakthrough of this period is arguably the synthetic production of precursors to artemisinin, an antimalarial drug produced naturally by the plant *Artemisia annua*, in the yeast *Saccharomyces cerevisiae* [Martin et al., 2003; Ro et al., 2006]. This progress has significantly improved our quantitative understanding of biological systems, with respect to the characterisation and availability of circuits and modules, and has revealed the engineering principles required of a proficient design strategy.

The current transition into the second wave of synthetic biology promises to prioritise the development of systems-level circuitry through the assembly of characterised modules [Purnick and Weiss, 2009]. It is here that many of the challenges and potential pitfalls faced by the community will have the greatest effect. For example, issues relating to standardisation will continue to resurface as synthetic biologists look to develop the modularity of devices across a multitude of cellular and environmental contexts. Intercellular, intracellular and extracellular properties

all present a daunting task concerning compatibility, particularly as the complexity of synthetic systems increases. Noise is another potentially problematic factor that directly affects the reliability and predictability of systems. Understanding noise and developing new methods for overcoming the associated ramifications will represent major progress towards realising expected system dynamics [Purnick and Weiss, 2009]. Epigenetics, the change in heritable information that occurs at a transcriptional level in response to environmental factors, without any alteration to the nucleotide sequence itself, also raises issues when considering how an engineered module might be suitably reset following an epigenetic event that compromises its desired functionality [Zheng et al., 2008].

The fundamental distinction between biological and physical systems regarding noise and other internal and external perturbations implies that it is entirely feasible that the methods required to realise the true potential of synthetic biology will inevitably be equally distinctive [Church et al., 2014]. Well established engineering principles have contributed to the *status quo*, however, their adequacy will no doubt be severely scrutinised as we confront the most complex and ground-breaking applications in synthetic biology. As a result, the standardisation and development of the methodologies required to engineer next-generation synthetic biological systems may need to take equal standing with device characterisation and modularity as the top priorities in the field. Indeed, biology may require its own specific branch of engineering altogether; concerns regarding module compatibility might only be solved by switching the focus of engineering away from individual circuits and towards the environments that pose the greatest threat to reliable deployment. A synthetic cell organelle capable of overseeing highly modular system operations across all cells and organisms could therefore be a defining goal for the field.

This being said, the number of novel circuits being published continues to rise in light of concerted efforts to overcome the aforementioned challenges [Cameron et al., 2014]. Further work on the antimalarial drug artemisinin has led to increased efficiency of its commercial production through a redeveloped artemisinic acid biosynthetic pathway [Paddon et al., 2013]. Biofuel production in *E. coli* has been made possible by adapting the cell’s natural amino acid biosynthetic pathway for alcohol synthesis [Atsumi et al., 2008]. An alternative approach to biofuel synthesis employs a dynamic sensor-regulator system (DSRS) to regulate biodiesel expression via control of a transcription factor, that is, a protein that controls the rate of transcription of mRNA from DNA by binding to specific DNA sequences [Zhang et al., 2012]. A host of other engineered microbial devices have also emerged; synthetic quorum sensing circuitry has enabled *E. coli* to detect tumour microenviron-



ments and respond by invading cancer cells [Anderson et al., 2006], edge-detection circuitry can perform computations of light images [Tabor et al., 2009] and bacteriophage (bacteria-infecting viruses) have been reprogrammed to target and degrade bacterial biofilms that are prominent in the development of diseases and exhibit antimicrobial resistance [Lu and Collins, 2007]. The scope of RNA-based devices has also expanded, with microbial kill switches that demonstrate improved leakage minimisation and modularity [Callura et al., 2010], and logic circuits that are capable of performing multiple Boolean operations governing the computation of gene expression and, in turn, cell function [Win and Smolke, 2008].

Although the range of new circuitry and their applications is highly diverse, the majority of existing systems are reliant, at least in part, on the detection and/or storage of information in eliciting an engineered response [Alon, 2006]. The storage of information, in particular, is crucial to the maintenance of reliable functionality and highlights the general importance of cellular memory. If bistability is exhibited by a transcriptional response, then an ‘on’ or ‘off’ genetic state is produced. The inheritance of this state, through DNA replication and cell division, provides a lasting memory of the response [Burrill and Silver, 2010]. The study of natural biological memory began over 50 years ago when French biologists François Jacob and Jacques Monod proposed a qualitative description of the link between cellular memory and transcriptional regulation [Burrill and Silver, 2010]. Their experimental work on the activation and repression of enzymatic lactose metabolism in *E. coli* that became the ‘operon model’ won them the 1965 Nobel Prize in Physiology or Medicine with their colleague André Lwoff [Morange, 2010]. This conceptual breakthrough transformed perceptions regarding how the array of distinct tissues arising in multicellular organisms express different sets of genes, despite the fact that all cells contain the same heritable genetic material [Morange, 2013].

The *lac* operon has since become the foremost example of microbial gene regulation, by virtue of the research of Jacob, Monod and Lwoff, enabling *E. coli* to metabolise lactose in the absence of its primary carbon source, glucose. In response to the availability of environmental resources, *E. coli* switches to expression of the *lac* operon genes which produces  $\beta$ -galactosidase, an enzyme capable of breaking down lactose into glucose and galactose [Morange, 2013]. Memory has been identified in many other natural environments: the bacteriophage  $\lambda$  is known to adopt two distinct life cycles on infection of a host bacterium, namely, a lysogenic state which allows the phage to live dormant within the host or a lytic state in which the host is destroyed and the viral progeny released [Ptashne, 2004]. As an example of cellular memory, eliciting a lasting response to a transient stimulus,  $\lambda$  is particularly

efficient: the lysogenic phage can remain dormant indefinitely until induction causes lytic growth across the entire population [Ptashne, 2006]. The immune and nervous systems also have the capacity for the retention of large quantities of information in order to remember, and act upon, new and old unfamiliar antigens or to manage the trillions of synaptic transmission events that occur in the brain [Burrill and Silver, 2010]. These systems present a clear motivation for engineering synthetic circuits that possess memory capable of storing information in the form of genetic states and highly stable maintenance of dynamical responses.

Synthetic transcriptional memory circuits that apply the core principles introduced by the toggle switch have demonstrated novel functionality within bacteria [Kobayashi et al., 2004] and other cellular contexts such as yeast [Becskei et al., 2001; Ajo-Franklin et al., 2007] and mammalian cells [Tigges et al., 2009; Kramer et al., 2003, 2004b]. Engineered *E. coli* can retain memory of DNA damage, demonstrating everlasting phenotypic changes [Kobayashi et al., 2004]. A genetic switch has been constructed in *S. cerevisiae*, providing an autocatalytic bistable cellular response in the presence of continuous stimulation [Becskei et al., 2001] and bistability in yeast is also demonstrated through positive feedback [Ajo-Franklin et al., 2007]. The toggle switch presented in Chinese hamster ovary (CHO) cells [Tigges et al., 2009] is based largely on its bacterial counterpart [Kobayashi et al., 2004] and realises the potential for epigenetic transgene control in mammals. Such genetic switches, capable of producing and sustaining two or more distinct genetic states upon transient induction, demonstrate behaviour analogous to that of transistors in electronic circuitry and thus have considerable potential in engineering cellular memory.

Transient operation of memory devices is an important criterion since continuous exposure to the relevant inducer is likely to be undesirable, especially in the realm of drug-inducible systems. A pioneering synthetic event-counter [Friedland et al., 2009] demonstrates how *E. coli* can be programmed to count up to three induction events through two distinct toggle switch circuit designs. The first design, termed the riboregulated transcriptional cascade (RTC), utilises a chain of transcriptional elements to record the expression of specific proteins in response to pulses of arabinose. The second design, termed the DNA invertase cascade (DIC), utilises proteins capable of specific binding to DNA recognition sites resulting in the subsequent rearrangement of intermediate genetic sequences, a process called DNA recombination, which provide inducible gene expression that is hard-wired into the cellular DNA. Both approaches are viable for engineering cellular memory [Friedland et al., 2009] and thus present synthetic biologists with two effective modules for incorporating memory into higher-level systems. For example, engineering human

immune cells could dramatically improve the efficacy of cancer therapies. T cells are a type of human immune cell or lymphocyte that are responsible for immune responses to tumour development. However, tumour cells have become increasingly adept at evading immune responses, and thus adoptive immunotherapy treatments are in development to amplify the efficacy of T cell responses. Current strategies involve removing T cells from the blood of a patient then making the appropriate genetic modifications to target the specific tumour microenvironment, before finally reintroducing the cells into the patient [Kershaw et al., 2013]. However, off-target effects of the engineered T cells have led to clinical toxicities and while safety ‘kill’ switches can nullify such problems, this also terminates the therapy. Synthetic memory-based circuitry is capable of introducing sophisticated ‘pause’ switch dynamics, facilitating inducible, temporal control of the engineered immune response. Characterising the cellular profile of populations of cancer cells and engineering an immune response to such phenomena would inhibit the development of cancerous tumours and other diseases that proliferate in a similar fashion [Burrill and Silver, 2010].

## 2.2 An introduction to mathematical modelling

A key tool in synthetic biology is mathematical modelling. By building a mathematical model of a system of interest, it is easier to develop an understanding of the relevant biology, and it is also relatively quick and simple to examine certain aspects of the system through model development. The efficacy of mathematical modelling is, however, directly proportional to the accuracy with which the biological system in question can be represented mathematically.

In the majority of cases, it is sufficient to model biological systems using ordinary differential equations (ODEs) since most systems consist of molecular reactions to which the law of mass action can be applied. This approach can capture the dynamics of a system over time with sufficient accuracy, but it does have limitations. Being entirely deterministic, ODEs fail to represent spatial components and do not account for any degree of stochasticity. When spatio-temporal detail is required from a model, solutions can be found using partial differential equations (PDEs). PDEs capture not just dynamic trends but also allow analysis of variables with respect to changes in their location. However, this added detail requires significant mathematical investment which may render the analysis unjustifiably complex. An alternative to PDEs is agent-based modelling (ABM) which is also capable of accounting for spacial properties by analysing the system components as

their populations develop and simulating the interactions that arise within a virtual environment. Accounting for stochasticity can be crucial in gaining an accurate representation of a system and may require approaches such as stochastic differential equations (SDEs) or hybrid systems. Modelling challenges for synthetic biologists usually involve finding the most accurate mathematical representation of the given biological system and deciding on a modelling strategy that presents good value for mathematical investment [Voit, 2013].

The output of a mathematical model is dependent on the associated model parameters which, in the case of biological systems, often represent the rates of the associated molecular reactions. Tuning sets of parameter values governing a mathematical model will vary its outputs in ways that can align simulations to real world observations [Arpino et al., 2013]. Such observations often take the form of experimental data and a wide variety of analytical tools are available to ascertain the strength of the model in replicating a desired output. Model development is therefore an iterative process that tests adaptations made to a model in light of apparent discrepancies and continues until a number of conditions, including optimisation and biological plausibility, are met [Klipp et al., 2005].

### 2.2.1 Mathematical models in synthetic biology

Although biological experimentation will always provide the keenest insight into system dynamics, if performed correctly, characterisation of the relevant circuitry is ideally suited to mathematical modelling approaches. As an additional tool in the analysis of system dynamics, mathematical models can achieve highly accurate simulations and predictions of biological systems. The statistician George Box said that “*All models are wrong but some are useful*”, a sentiment that has been adopted widely in academic communities to encapsulate the *status quo*. At face value, this may appear to be a pessimistic outlook however it is the subset of useful models that have the potential to elucidate systems in biology, as well as numerous other research fields. It must be accepted that all models are wrong simply because the systems that they are designed to simulate are almost always so complex that it is impossible to account for all of the associated variables. Attempting to construct an exact model that simulates a given system perfectly is therefore futile; after all, by definition, a model is a representation and does not claim to be anything more than that. That said, significant qualitative, and often quantitative, system dynamics can be captured by mathematical models of varying complexity, making them very useful in synthetic biology.

Fundamentally, mathematics offers a highly rigorous framework with which

to model and analyse natural systems. For centuries mathematics has been viewed as the language of nature, forming the basis of multiple fundamental principles across the sciences. Applied mathematics, in particular, facilitated the development of calculus in response to progressive ideas in physics and has subsequently led to the establishment of many, if not all, of the mathematical branches comprising classical engineering disciplines. With regards to biology, mathematical applications have established numerous subject areas that constitute the field of mathematical biology in its own right. For example, exponential and logistic growth functions are intrinsic to the study of standard population dynamics and can simulate the behaviour of more intricate systems such as the oscillations observed in predator-prey populations. Analysis of infectious disease epidemics is also fundamentally based on an array of mathematical modelling approaches which ultimately enables health authorities to intervene effectively; selection of the most suitable approach is dependent upon the assumptions made with regard to the specific context of the problem at hand.

Mathematics is also fundamental to systems biology, facilitating the analysis of dynamical systems relating to gene regulation and many other intercellular and intracellular processes including cell-cell signalling and chemotaxis. Systems biology constitutes a large part of synthetic biology since dynamical systems analysis is entirely relevant and crucial to the research effort, with the added emphasis on novel circuit design setting synthetic biology apart. Mathematical frameworks offer an ideal platform for the generation of computer code in an age of ongoing expansion in computing power. Programming software has become prevalent in both academia and industry due to its speed and reliability. The language of computing was borne out of mathematics and subsequently presents the ideal environment for the deployment and development of numerical tasks. For example, numerical analysis provides an array of algorithmic methods that facilitate the computer-aided calculation of model outputs.

One of the main benefits of mathematical modelling in synthetic biology is the relatively short time frames required to produce simulations and other outputs of interest. Whilst experimentation provides the most trusted evidence of dynamical behaviour, procedures necessarily require long time frames for completion and are invariably subject to human error. Even experiments that are performed successfully are influenced by inherent noise that can skew results and hence collating good data is time consuming, especially with the additional requirement that experiments be repeated multiple times to prevent the propagation of anomalous results. The *in silico* simulations that models provide can be thought of as computational

experiments that are performed on the order of seconds. Therefore, practical experimentation can be directed by computational results and thus development times are greatly reduced through an optimised allocation of resources. Furthermore, certain aspects of system dynamics are particularly difficult to ascertain due to the limitations of current experimental procedures, but a useful model has the potential to give insights into any such aspect. This means that not only is the standard practice of experimental observation and measurement providing the necessary benchmarks with which to align model performance widely accepted, but also that the contrary is valid; model simulations and predictions are able to reveal any number of significant mechanistic properties that might require further experimental verification.

Of course, mathematical models are highly rigorous representations of their associated natural systems as long as they are derived correctly and are well informed. Model derivation typically utilises the law of mass action, a long standing method that translates the rate of change of the physical quantity of a particular entity into a mathematical framework. Models constructed in this way are referred to as deterministic and are commonplace in the literature [Elowitz and Leibler, 2000; Gardner et al., 2000; Tindall et al., 2012; Bonnet et al., 2012; Hancock et al., 2015] however, many other derivation techniques are valid in the appropriate context and can provide equally insightful results. The reaction networks that inform model derivation are formulated based on biological knowledge and hence the quality of a model is directly proportional to the quality of the information available in biological literature. A validated mathematical model that exhibits useful, user-defined outputs can be adapted to emulate any number of experimental conditions and can be disseminated easily without loss of information. Model validation is largely dependent on, but not limited to, the analysis of the fitness of simulations compared to experimental observations. Hence, the collection and dissemination of well established experimental data is vital in predictive modelling efforts and illustrates further the importance of the interdisciplinary nature of the synthetic biology community.

### **2.2.2 The trade-off between mechanistic and black box modelling approaches**

Selection of a modelling strategy is dependent upon the nature of the associated biological system and the desired output, and there exists numerous mathematical frameworks that facilitate a wide variety of model analysis techniques [Voit, 2013]. Arguably, the more important choice relates to the degree of accuracy required from the model. The trade-off between simplicity and accuracy is at the core

off all modelling investigations and useful results are unlikely to be obtained without establishing an effective compromise. Generally, accounting for minimal detail decreases mathematical investment whereas increased complexity requires greater investment. The former case is often referred to as black box modelling; black box models are typically over-simplified in order to provide generic, qualitative outputs and are so named due to their lack of mechanistic detail which boils down to a case of input-output with little knowledge or clarity regarding the relevant system structure. The latter case is often referred to mechanistic modelling; mechanistic models account for specific biological mechanisms and other structural details with a view to providing quantitative, physically valid outputs. Note that a spectrum of mechanistic complexity exists which places black box models at one extreme and white box models (white box models will be referred to as mechanistic models for the purposes of this work) at the other extreme with grey box models referring to any intermediate modelling approach that contains a mixture of elements.

The distinction between black box and mechanistic modelling can be illustrated by considering the enzyme-substrate interaction model that forms the basis of Michaelis-Menten enzyme kinetics:



where  $E$  is the enzyme,  $S$  is the substrate,  $C$  is the enzyme-substrate complex formed in the reaction and  $P$  is the product formed by the reaction;  $k_1$  and  $k_{-1}$  represent the forward and backward reaction rates of the reversible complex formation respectively and  $k_2$  represents the reaction rate of the irreversible product formation. In a black box modelling approach, we over-simplify these interactions by neglecting the formation of the enzyme-substrate complex and, instead, consider a single reaction that results in product formation via enzyme-substrate binding:



where  $k_3$  represents the reaction rate of the irreversible product formation and encapsulates all three reaction rates from (2.1). The interactions denoted by (2.2) and (2.1) can therefore inform the derivation of black box and mechanistic models of enzyme kinetics respectively. The black box model exhibits the same dynamics for both the enzyme and the substrate as their concentrations decrease towards zero over time; the product is shown to increase in concentration as expected (Figure 2.4A). In contrast, the mechanistic model includes the added dynamics of the enzyme-substrate complex and exhibits subtly different outputs overall (Figure 2.4B). The

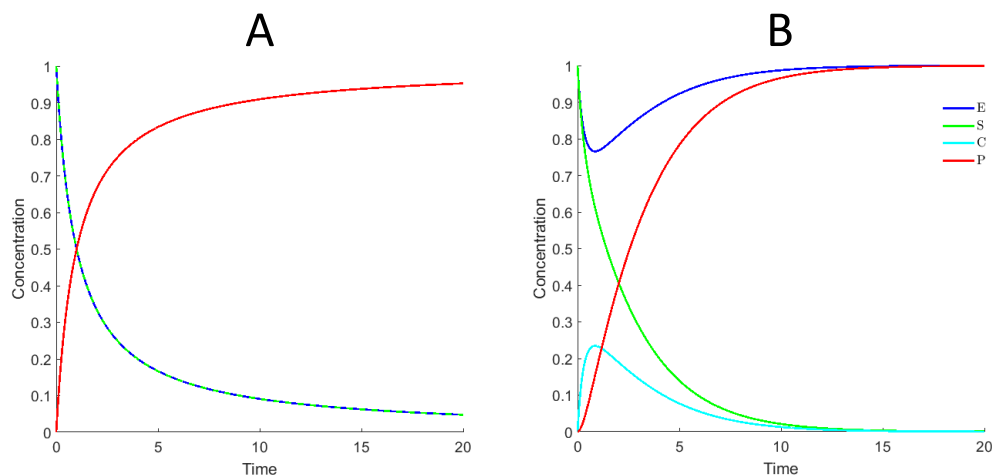


Figure 2.4: Comparison of black box and mechanistic enzyme kinetics model outputs. A) The black box model accounts for product,  $P$ , formation by virtue of enzyme,  $E$ , and substrate,  $S$ , binding interactions. B) The mechanistic model accounts for additional detail regarding the formation of the intermediate complex,  $C$ . All parameters and initial concentrations of enzyme and substrate used in simulations set equal to 1; initial concentrations of complex and product set equal to 0.

enzyme dynamics in particular are significantly different since the black box model does not account for enzyme dissociation from the complex. Although the substrate and product dynamics are qualitatively similar for both models, there are clear distinctions that highlight the consequences of over-simplification.

Comparisons of the black box and mechanistic models demonstrate the underlying system dynamics that can be overlooked through generalisation. This example is, however, particularly simplistic even in the case of our mechanistic model considering the intricacy that can potentially arise in biological systems. The mechanistic model consists of one more biological entity and two more reaction rates than the black box model which presents a negligible increase in mathematical investment. Numerous mathematical modelling techniques are required to produce the simulations in Figure 2.4, the details of which are given in the following sections. Note that these simulations require information regarding the parameter values corresponding to the associated reaction rates as well as the initial concentrations of the associated biological entities. The parameter values,  $k_1$ ,  $k_{-1}$ ,  $k_2$  and  $k_3$  in this example are all arbitrarily set equal to 1 for comparable simulations. The initial conditions are selected given the context of the problem; only the enzyme and substrate are present at time zero and are therefore given arbitrary initial concentrations of 1 whereas the



complex and product are formed over time and are not present initially, resulting in initial concentrations of 0.

It is mechanistic modelling approaches that are adopted throughout this thesis in order to construct the most structurally detailed biological models possible that can be validated quantitatively, and thus provide physically valid outputs. The systems-level understanding and experimental data available to us is particularly well founded by virtue of multiple collaborative efforts and extensive literature mining. Despite the substantial increase in mathematical investment presented by the large, complex models that we aim to produce, we are confident that rigorous mathematical analysis poses a manageable obstacle, the cost of which is outweighed by the insight gained from our results.

## 2.3 Mathematical modelling techniques

### 2.3.1 Ordinary differential equations and initial value problems

Mathematical models of physical systems require the appropriate mathematical framework for describing specific processes and interactions. For example, if two variables,  $x$  and  $y$ , are directly proportional to each other we write,

$$y \propto x. \quad (2.3)$$

This describes the relationship whereby an increase/decrease in the variable  $x$  results in an increase/decrease in the variable  $y$  and vice versa. There is, however, a lack of information regarding the magnitude of the increase in  $y$  for a given increase in  $x$ . We therefore transform (2.3) into the following mathematical equation,

$$y = kx, \quad (2.4)$$

where  $k$  is a constant, known as the coefficient of proportionality, and hence the ratio of  $x$  and  $y$  is equal to this constant value,

$$k = \frac{y}{x}, \quad (2.5)$$

for all non-zero values of  $x$  and  $y$ . Figure 2.5 illustrates how the relationship between  $x$  and  $y$  is influenced by  $k$ . When  $k = 1$ ,  $y = x$  and thus  $y$  bisects the positive quadrant of the axes through the origin. When the value of  $k$  is increased i.e.  $k = 5$ , the magnitude of  $y$  is increased fivefold for every  $x$  value thus producing a steeper line. When the value of  $k$  is decreased i.e.  $k = 0.2$ , the magnitude of  $y$  is decreased

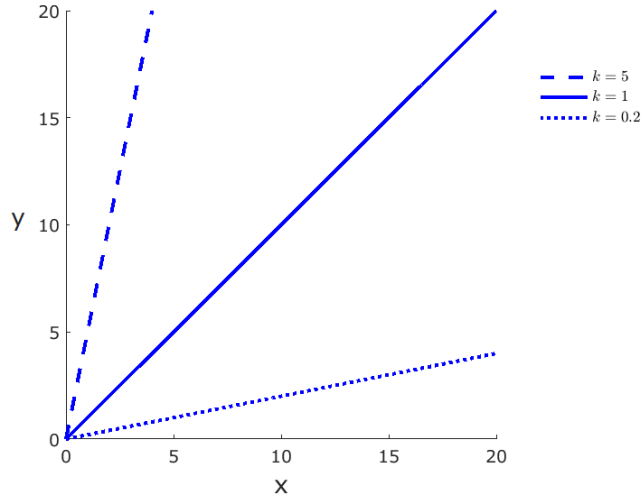


Figure 2.5: Comparisons of varied  $k$  values in the linear function  $y = kx$ . The gradient of the line increases for larger values of  $k$  and decreases for smaller values of  $k$ .

fivefold for every  $x$  value thus producing a line with decreased steepness. This direct proportionality is exhibited by the relationship between the diameter of a circle and its circumference, that is,

$$C = \pi d, \quad (2.6)$$

where  $C$  and  $d$  denote the circumference and diameter of a circle respectively and the coefficient of proportionality is equal to the irrational constant  $\pi$ . Proportionality gives rise to simple linear relationships such as (2.6) however, the same principles can also be applied to more intricate and complex relationships. Mathematical models are often formulated in light of information regarding the rate of change in the relevant variables over time. In order to describe the rate of change over time mathematically for a given variable  $y$ , we take the derivative with respect to time,  $t$ ,

$$\frac{dy}{dt}, \quad (2.7)$$

where  $dy$  and  $dt$  denote the change in  $y$  and the change in  $t$  respectively. Using this framework, we are able to construct mathematical models of systems in which the rate of change in the output variable of interest is proportional to the variable itself. For example, consider a population growth model whereby the rate of change of the

population at any given time is directly proportional to that population,

$$\frac{dP}{dt} \propto P, \quad (2.8)$$

where  $P$  and  $t$  denote population and time respectively. The variable  $P$  on the right hand side is positive, indicating population increase or growth; negative terms are used to model the behaviour of negative growth or decay in such systems. This relationship is transformed into a mathematical equation in the same way as (2.3), that is,

$$\frac{dP}{dt} = kP, \quad (2.9)$$

where the constant  $k$  is a newly derived coefficient of proportionality. Both  $P$  and its derivative, or differential, appear in (2.9) and hence we refer to this equation as an ordinary differential equation (ODE). In order to determine the function that describes the growth of this population over time, we must solve this ODE and therefore obtain a function for  $P$  only. Solving ODEs can involve many different calculus-based methods and techniques depending on the nature of the given equation. This particular case is sufficiently straightforward to employ the method of separation of variables as follows:

$$\begin{aligned} & \frac{dP}{dt} = kP, \\ \Rightarrow & \frac{1}{P} dP = k dt, \\ \Rightarrow & \int \frac{1}{P} dP = \int k dt, \\ \Rightarrow & \ln P = kt + c, \\ \Rightarrow & P = \exp(kt + c), \\ \Rightarrow & P = \exp(kt) \exp(c), \\ \therefore & P = A \exp(kt), \end{aligned}$$

where  $c$  is the constant of integration and  $A = \exp(c)$  is also constant. Assuming that we have sufficient knowledge of the system, we can make an appropriate estimate for the value of the constant parameter  $k$ . Taking  $k = 0.5$  we have,

$$P = A \exp(0.5t). \quad (2.10)$$

At this stage, (2.10) is a general solution to the ODE by virtue of the fact that, although we have determined the function that describes the evolution of the population over time, the function can exhibit an infinite number of solutions dependent on the value of the constant  $A$ . In order to determine the exact solution to the ODE, we require additional information regarding the size of the population at a given time point. This is known as an initial condition since the information is given for the case when time is zero and consequently, together with the original ODE (2.9), describes an initial value problem (IVP). In this example, consider the initial condition that the population is equal to ten when time is zero,  $P = 10$  at  $t = 0$ , then,

$$\begin{aligned} &10 = A \exp(0), \\ \implies &A = 10, \\ \therefore &P = 10 \exp(0.5t). \end{aligned} \tag{2.11}$$

The function (2.11) is the exact solution to the IVP, exhibiting a single trajectory describing the growth of the population (Figure 2.6). The function can be used

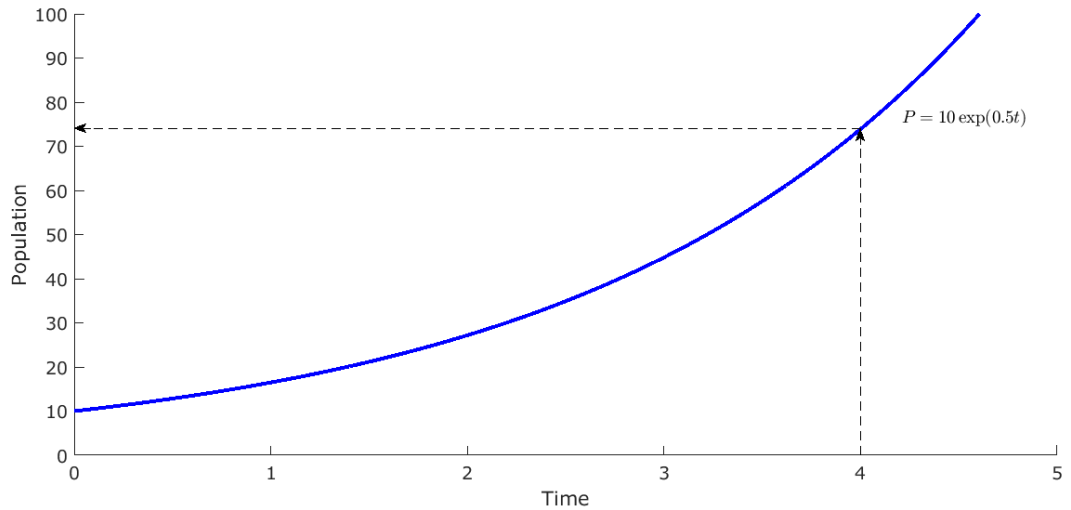


Figure 2.6: Exact solution to the IVP described by (2.9) and the initial condition  $P = 10$  at  $t = 0$  ( $P(0) = 10$ ). The dashed arrows depict the graphical prediction of the population size at  $t = 4$ .

to predict the size of the population at future time points by simply evaluating the function at the time point of interest. For example, if we want to know the projected population at four time points after population growth was initiated, we determine

$P$  at  $t = 4$ ,

$$\begin{aligned} & \Rightarrow P = 10 \exp(0.5 \times 4), \\ & \therefore P = 10 \exp(2), \\ & \therefore P = 73.9. \quad (1 \text{ d.p.}) \end{aligned}$$

Similarly, if we want the time taken for the population to reach a particular size then we can substitute the desired population size into (2.11) and solve for  $t$  instead. Note that the choice of the constant parameter  $k$  in (2.10) was made assuming sufficient knowledge of the system. However, in practice, it is common that the relevant model parameters are not well established. This has implications regarding the interpretation and plausibility of solutions to IVPs that will be covered in further detail in the remainder of this chapter.

### 2.3.2 The law of mass action

The application of mass action kinetics to systems of biological reactions is fundamental to the construction of deterministic mathematical models. The law states that the rate of a chemical reaction is directly proportional to the product of the reactants. This facilitates the derivation of a system of ODEs from the relevant system of biochemical equations representing the dynamical interactions of a biological reaction network. Each ODE in the model describes the dynamical evolution of the corresponding biological entity with respect to time. For example, consider a basic biological process in which A reacts with B to form C at a rate  $k$ :



where A, B and C represent three distinct biological entities. Application of the law of mass action enables us to derive the rate of change in molecular concentration for each of these entities; molecular concentration being the standard dimensionality in mathematical modelling efforts. Firstly, the rate of change in concentration of A is proportional to the product of A and B and is also negative since A is depleted in the reaction:

$$\frac{d[A]}{dt} = -k[A][B], \quad (2.13)$$

where the reaction rate  $k$  equates to the coefficient of proportionality and the square bracket notation is used to denote concentration. Secondly, B plays an equivalent role in the reaction to A and therefore we derive an equivalent ODE for the rate of

change in concentration of B:

$$\frac{d[B]}{dt} = -k[A][B]. \quad (2.14)$$

Thirdly, the rate of change in concentration of C is proportional to the product of A and B, but is positive since C is accumulated in the reaction:

$$\frac{d[C]}{dt} = k[A][B]. \quad (2.15)$$

Having derived the full mathematical model for this basic biological reaction, we have identified the reaction rate  $k$  as the sole model parameter. The magnitude of this parameter will determine the dynamical responses of the three dependent variables (biological entities). Since this model encompasses just one reaction it is no surprise that the same term appears in each model ODE, albeit with varying sign. In general, biological networks of interest consist of many different molecular interactions that translate to ODEs with many different terms, the extent to which is directly influential in selecting methods of ascertaining the appropriate solutions.

### 2.3.3 Explicit and numerical solutions to initial value problems

Simulating biological system dynamics *in silico* involves solving the associated model ODEs. The previous example model (2.13)-(2.15) is sufficiently simple to solve explicitly via calculus, subject to the appropriate initial conditions. At the start of the reaction, the reactants A and B are present whereas C is yet to be produced and hence we define the following initial conditions, dropping the square bracket notation for convenience:

$$\text{at } t = 0 : \quad A = A_0 > 0, \quad B = B_0 > 0, \quad C = C_0 = 0, \quad (2.16)$$

where  $A_0$ ,  $B_0$  and  $C_0$  are constants representing the initial concentrations of A, B and C respectively. Subtracting (2.14) from (2.13) enables the derivation of a conservation relation on A and B (we assume that all matter is conserved within the system, with the specific quantities described by the initial concentrations):

$$\begin{aligned} & \frac{dA}{dt} - \frac{dB}{dt} = -kAB - (-kAB) = 0, \\ \implies & \int \frac{dA}{dt} - \frac{dB}{dt} dt = \int 0 dt, \\ \implies & A - B = c, \end{aligned}$$

$$\begin{aligned}
&\Rightarrow c = A_0 - B_0, \quad (\text{at } t = 0) \\
\therefore A - B &= A_0 - B_0,
\end{aligned} \tag{2.17}$$

where  $c$  is the constant of integration and  $t$  is the independent time variable. The conservation relation (2.17) takes two forms dependent on the nature of the initial concentrations of the reactants i.e.

$$\begin{cases} A = B, & (A_0 = B_0) \\ A = B + A_0 - B_0. & (A_0 \neq B_0) \end{cases} \tag{2.18}$$

In the case where the two initial concentrations are equal, (2.13) becomes:

$$\frac{dA}{dt} = -kA^2,$$

and hence,

$$\begin{aligned}
&\frac{1}{A^2} dA = -k dt, \\
\Rightarrow \int \frac{1}{A^2} dA &= \int -k dt, \\
\Rightarrow -\frac{1}{A} &= -kt + c, \\
\Rightarrow c &= -\frac{1}{A_0}, \quad (\text{at } t = 0) \\
\Rightarrow -\frac{1}{A} &= -kt - \frac{1}{A_0}, \\
\therefore A &= \frac{1}{kt + \frac{1}{A_0}}.
\end{aligned} \tag{2.19}$$

Substituting (2.19) into (2.15) gives

$$\begin{aligned}
&\frac{dC}{dt} = \frac{k}{(kt + \frac{1}{A_0})^2}, \\
\Rightarrow \int dC &= \int \frac{k}{(kt + \frac{1}{A_0})^2} dt, \\
\Rightarrow C &= -\frac{1}{(kt + \frac{1}{A_0})} + c, \\
\Rightarrow c &= C_0 + \frac{1}{(0 + \frac{1}{A_0})} = A_0, \quad (\text{at } t = 0) \\
\therefore C &= A_0 - \frac{1}{(kt + \frac{1}{A_0})}.
\end{aligned}$$

Solving the system for the case where the two initial concentrations of the reactants are not equal in a similar manner provides the full explicit solution for  $C$ :

$$\begin{cases} C = A_0 - \frac{1}{kt + \frac{1}{A_0}}, & (A_0 = B_0) \\ C = \frac{\exp((B_0 - A_0)kt) - 1}{\frac{1}{A_0} \exp((B_0 - A_0)kt) - \frac{1}{B_0}}. & (A_0 \neq B_0) \end{cases} \quad (2.20)$$

Generally, mathematical models of large and complex biological systems rarely exhibit explicit solutions since the integration step of the necessary calculations is often intractable using standard calculus. In these cases, numerical programming software such as MATLAB is used to compute model outputs. Such software employs an array of numerical methods in order to obtain accurate approximations to integrals that give rise to the relevant solution trajectories. An integral is defined as the area under the curve described by a given function, bounded by an interval in the domain of the function. A standard approach to approximating integrals is the trapezium rule which calculates the area of one or more trapeziums of similar dimensions to the bounded area under the curve. The area of a trapezium is given by,

$$A = h \left( \frac{a + b}{2} \right) \quad (2.21)$$

where  $a$  and  $b$  are the lengths of the two parallel sides of the trapezium and  $h$  is the height, the distance between these two sides (Figure 2.7A). Hence, for a function

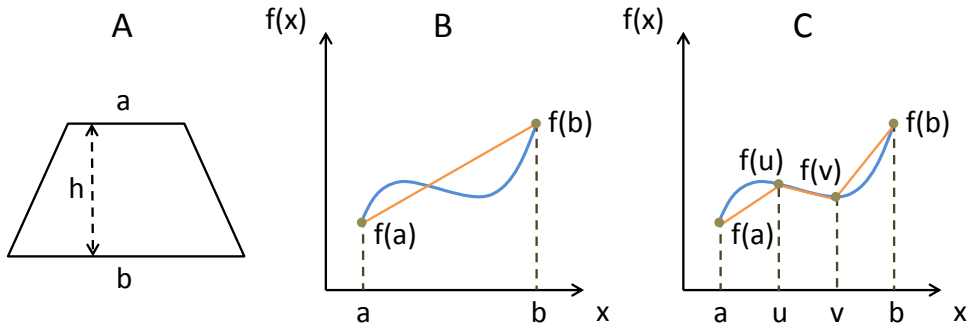


Figure 2.7: The trapezium rule. A) Diagram of a standard trapezium and the necessary labelled dimensions required to calculate its area. B) Graph depicting the approximation of an integral of the function  $f(x)$  using the trapezium rule with one trapezium ( $n = 1$ ). C) Graph depicting the approximation of an integral of the function  $f(x)$  using the trapezium rule with three trapeziums ( $n = 3$ ). Functions and trapezium edges are plotted using blue and orange lines respectively.



$f(x)$ , the integral of  $f(x)$  in an interval of its domain  $[a, b]$  can be approximated by the area of the trapezium described by the length of the interval,  $b - a$ , the function evaluations at the endpoints of this interval,  $f(a)$  and  $f(b)$ , and the line connecting these two points (Figure 2.7B). The area of this trapezium is therefore given by,

$$A = (b - a) \left( \frac{f(a) + f(b)}{2} \right), \quad (2.22)$$

where the length of the interval  $[a, b]$  is  $(b - a)$  corresponding to  $h$  and, consequently, (2.22) gives an approximation to the integral of  $f(x)$  in this interval,

$$\int_a^b f(x) dx \approx (b - a) \left( \frac{f(a) + f(b)}{2} \right). \quad (2.23)$$

Note that this approximation arises from the straight line connecting  $f(a)$  and  $f(b)$  which causes either an overestimate or underestimate of the area under the curve depending on the nature of the function in the given interval (Figure 2.7B). The accuracy of the approximation can be improved by dividing the interval into multiple trapeziums of equal width and summing the areas of the individuals (Figure 2.7C). For example, if we divide the interval  $[a, b]$  into three equal parts by introducing intermediate points  $u$  and  $v$  we can describe the area of three trapeziums using the function evaluations at these points ( $f(a)$ ,  $f(u)$ ,  $f(v)$  and  $f(b)$ ) and the distances between them, such that,

$$\begin{aligned} \int_a^b f(x) dx \approx & (u - a) \left( \frac{f(a) + f(u)}{2} \right) + (v - u) \left( \frac{f(u) + f(v)}{2} \right) + \dots \\ & (b - v) \left( \frac{f(v) + f(b)}{2} \right), \end{aligned}$$

where the heights of the trapeziums are equal, each being one third of the length of the whole interval, and hence,

$$\begin{aligned} \int_a^b f(x) dx \approx & \frac{(b - a)}{3} \left( \left( \frac{f(a) + f(u)}{2} \right) + \left( \frac{f(u) + f(v)}{2} \right) + \left( \frac{f(v) + f(b)}{2} \right) \right), \\ \approx & \frac{(b - a)}{3} \left( \frac{f(a)}{2} + \frac{f(u)}{2} + \frac{f(u)}{2} + \frac{f(v)}{2} + \frac{f(v)}{2} + \frac{f(b)}{2} \right), \\ \approx & \frac{(b - a)}{3} \left( \frac{f(a)}{2} + f(u) + f(v) + \frac{f(b)}{2} \right), \\ \approx & \frac{(b - a)}{6} (f(a) + 2f(u) + 2f(v) + f(b)). \end{aligned}$$

The improvement in accuracy of the approximation gained by increasing the number of trapeziums from one to three can be seen by comparing the area between the orange and blue lines in Figure 2.7B and Figure 2.7C. Approximating the integral by a number of trapeziums  $n$  gives the general formula for the trapezium rule,

$$\int_a^b f(x) dx \approx \frac{h}{2n} (f(a_1) + 2f(a_2) + 2f(a_3) + \cdots + 2f(a_n) + f(a_{n+1})), \quad (2.24)$$

in the interval  $[a, b]$  divided into  $n$  equal parts such that  $a = a_1 < a_2 < a_3 < \cdots < a_{n+1} = b$ . The most accurate approximations are achieved for large  $n$  values and hence, although this results in a large number of calculations that would quickly become unmanageable by hand, this method is ideally suited to computational analysis whereby vast numbers of calculations can be performed in very short time frames. In practice, numerical integration methods employed computationally are performed iteratively, in the form of algorithms, since this provides a large number of simple calculations each of which is dependent on the prior solution. The trapezium rule forms the basis of an equivalent iterative method known as the second order Runge-Kutta method (RK2). Runge-Kutta methods are a family of iterative numerical integration techniques whose accuracy improves with increased order of the relevant equations. The most commonly used method in computational mathematical programming is the fourth order Runge-Kutta method (RK4). Consider an IVP:

$$\frac{dy}{dt} = f(t, y), \quad f(t_0) = y_0, \quad (2.25)$$

defined for a function,  $y(t)$ , where  $t$  is the independent time variable, as is the standard for modelling biological systems. In order to recover the function  $y$  from the derivative in (2.25) we can integrate both sides:

$$\begin{aligned} \int \frac{dy}{dt} dt &= \int f(t, y) dt, \\ \implies y &= \int f(t, y) dt. \end{aligned}$$

We therefore require the integral of the function  $f(t, y)$  and hence the solution to (2.25) can be approximated numerically via the RK4 equations:

$$\begin{aligned} y_{n+1}^* &= y_n^* + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4), \\ t_{n+1} &= t_n + h, \end{aligned}$$

where  $y_n^*$  is the approximate solution for  $y$  at time  $t_n$  for  $n = 0, 1, 2, \dots$ ;  $h$  is the step size and we have:

$$\begin{aligned}k_1 &= f(t_n, y_n^*), \\k_2 &= f(t_n + \frac{h}{2}, y_n^* + \frac{h}{2}k_1), \\k_3 &= f(t_n + \frac{h}{2}, y_n^* + \frac{h}{2}k_2), \\k_4 &= f(t_n + h, y_n^* + hk_3),\end{aligned}$$

where  $k_1$  gives the derivative of the function at the start point of the interval  $h(t_n)$ ,  $k_2$  gives the derivative of the function at the midpoint of the interval  $h(t_n + \frac{h}{2})$  using  $k_1$ ,  $k_3$  gives a second calculation of the derivative of the function at the midpoint using  $k_2$  and  $k_4$  gives the derivative of the function at the endpoint of the interval  $h(t_n + h)$ . This method allows the approximate solution at the next time point,  $y_{n+1}^*$ , to be calculated using the known information regarding the approximate solution at the previous time point,  $y_n^*$ . The four increments,  $k_{1,2,3,4}$  are four separate attempts to predict the next  $y^*$  value, each based on a different principle. The first increment,  $k_1$ , predicts the next approximate solution by virtue of the gradient at the start point; an assertion that in isolation forms the basis of the first order Runge-Kutta method (RK1), also known as the Euler method. The second increment,  $k_2$ , predicts the next approximate solution by virtue of the midpoint of the interval described by the step size  $h$  which, when utilised only with  $k_1$ , forms the basis of RK2. The third increment,  $k_3$ , predicts the next approximate solution by virtue of a secondary consideration of the gradient at the midpoint, using  $k_2$ . The fourth increment,  $k_4$ , predicts the next approximate solution by virtue of the gradient at the endpoint, using  $k_3$ . Finally, the overall calculation of  $y^*$  is determined based on a weighted average of all four increments, with the greater weight given to the increments at the midpoint; resulting from the derivation of the RK4 equations in a similar manner to that in the derivation of the trapezium rule (2.24).

The increased accuracy that RK4 provides by comparison to lower order methods arises from the increased number of increments subject to weighted averaging. The standard ODE solver in MATLAB is ode45 which employs a slight adaptation on RK4 known as the Dormand-Prince method (RKDP). This method has the advantage over standard RK4 due to the six increments that are used to calculate each approximate solution, as well as a more precise weighted average of these increments. RKDP is an adaptive Runge-Kutta method, capable of selecting adapted step sizes based on the error calculated at each step as the algorithm con-

verges. Standard ODE solvers such as ode45 provide sufficient accuracy for solutions to systems of one or more ODEs where the relevant parameter space is relatively small. In some cases, however, it is necessary that model parameters vary by several orders of magnitude, causing the rate of change of certain variables to evolve very quickly or slowly compared to others. These scenarios are referred to as stiff equations or models and standard numerical methods are unable to provide sufficient accuracy. MATLAB provides ODE solvers designed specifically for solving stiff systems, such as ode15s. In contrast to the numerical integration methods employed by non-stiff ODE solvers, the method employed by ode15s is the numerical differentiation formulas (NDFs). NDFs are a modification of backward differentiation formulas (BDFs). The backward difference calculations employed by NDFs facilitate the analysis of adaptive step-size problems and provide increased stability compared to standard BDFs, making them ideally suited to solving stiff equations. Note that non-stiff solvers such as ode45 are capable of solving stiff equations, but they take considerably longer due to the vast number of iterative steps required to establish a reasonable solution.

To illustrate the accuracy of numerical methods for solving systems of ODEs *in silico*, we compare both explicit and numerical solution trajectories for the example model (2.13)-(2.15) (Figure 2.8). No distinction can be made when comparing the explicit solutions for  $C$  (2.20) to the corresponding numerical solutions, for both sets of initial conditions. The effect of varying the value of the model parameter is also demonstrated; when  $k$  is given a relatively large value of 5 the solution reaches steady-state faster than when  $k$  is given a smaller value. Hence, the concentration of  $C$  is shown to increase faster for greater reaction rates and thus presents an expected result for this basic example. The numerical solutions in Figure 2.8 were computed in MATLAB using ode45 since the model is relatively small in size and contains just one parameter taking nominal values that does not present any stiffness.

#### 2.3.4 Model reduction

Conservation relations such as (2.17) provide simple linear relationships between model variables that, in addition to facilitating the calculation of explicit solutions, can enable model reduction. That is, the total number of model ODEs and parameters can be reduced which, in turn, simplifies the subsequent mathematical analysis of the model. For example, we can determine the linear relationship between the example model variables  $A$  and  $B$  by making  $A$  the subject of (2.17):

$$A = B + A_0 - B_0. \quad (2.26)$$

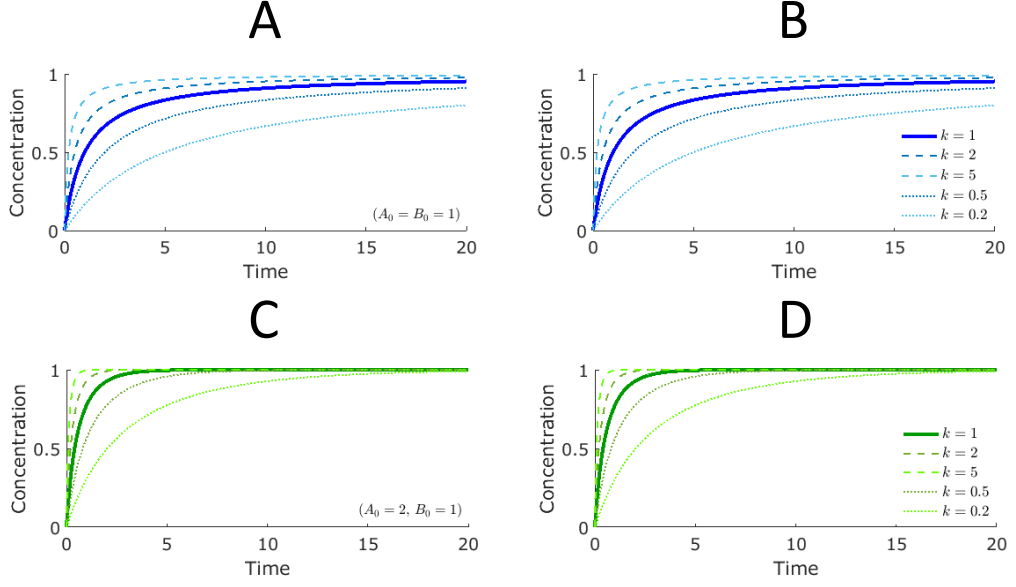


Figure 2.8: Comparison of explicit and numerical solutions to the example model. The explicit solution A) and the numerical solution B) are plotted for the initial condition  $A_0 = B_0 = 1$ . The explicit solution C) and the numerical solution D) are plotted for the initial condition  $A_0 = 2, B_0 = 1$ . Five distinct trajectories are plotted in each case for five values of the model parameter  $k$  such that  $k = \{0.2, 0.5, 1, 2, 5\}$ .

We are now able to eliminate the ODE associated with  $A$  and substitute (2.26) into the remaining model ODEs to yield the following reduced model:

$$\frac{dB}{dt} = -k(B + A_0 - B_0)B, \quad (2.27)$$

$$\frac{dC}{dt} = k(B + A_0 - B_0)B. \quad (2.28)$$

We can also sum (2.27) and (2.28) to determine a second conservation relation on  $B$  and  $C$ :

$$B = B_0 - C, \quad (2.29)$$

which allows us to eliminate the ODE associated with  $B$  and therefore yield the following fully reduced model:

$$\frac{dC}{dt} = k((B_0 - C) + A_0 - B_0)(B_0 - C), \quad (2.30)$$

$$= k(A_0 - C)(B_0 - C), \quad (2.31)$$

where the only dependent variable remaining in the system is  $C$  since  $A_0$  and  $B_0$  are constants. The numerical solution to (2.31) is identical to that of  $C$  in the full example model (Figure 2.9) and hence we have demonstrated that model reduction

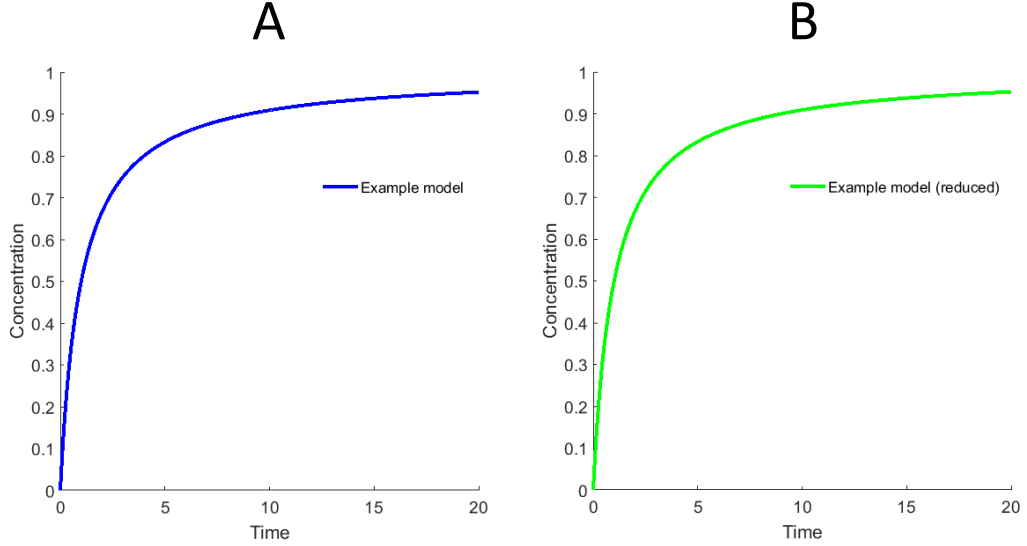


Figure 2.9: Numerical solutions of the example model. A) The solution trajectory for  $C$  arising from the numerical solution of the full example model consisting of three ODEs. B) The numerical solution of the reduced example model consisting of the single ODE governing the dynamical response of  $C$ .

retains mechanistic output. Model reduction can also be achieved through alternative methods such as the equilibrium and quasi-steady-state assumptions that assist in the reduction of the Michaelis-Menten enzyme kinetic model [Murray, 2002].

We have reduced the example model from three ODEs to just one, which is particularly beneficial since there is only one ODE to analyse and therefore the required mathematical investment has been minimised. Model analysis often involves the calculation of explicit solutions however, in the case of the example model, it is not apparent whether solving (2.31) explicitly via calculus would yield our known solution (2.20) with significantly less mathematical investment than the method outlined in section 2.3.3. The benefits of model reduction are more substantial for larger, mechanistic models consisting of many ODEs since, not only are the original mechanistic properties of the model preserved, but the reduction in dimensionality can alleviate the complexity of subsequent model analysis [Hancock et al., 2015].

The mechanistic models presented in this thesis are constructed for the purpose of simulation and prediction, with a focus on their ability to replicate experimental data and observations *in silico*. We are interested in verifying whether the

networks and system architectures, developed through literature mining and experimental collaboration, are suitable for deriving mathematical models that can provide quantitative predictions of the relevant system dynamics in order to facilitate novel circuit design. Hence, we do not attempt to reduce our mechanistic models and, instead, analyse the quality of model outputs as opposed to the underlying mathematical structure.

### 2.3.5 Non-dimensionalisation

Deriving and simulating mathematical models using the methods outlined in this section provide significant insights into the dynamical nature of the associated biological systems. The accuracy of model outputs is improved greatly by a strong knowledge of the relevant reaction rate constants since well established parameter values come associated with a specific dimensionality that is highly influential within the context of the model. However, it is usually the case that knowledge of a full parameter set for a given model is lacking due to difficulties recording the required measurements experimentally. In order to overcome such uncertainties, it is often beneficial to remove all dimensionality from a model so that direct mathematical comparisons can be made. This non-dimensionalisation is performed by writing each dependent and independent variable in terms of model parameters shown to possess the required dimensionality.

Considering the basic example model, we confirm the dimensionality of the dependent and independent variables:

$$[A] = [B] = [C] = \frac{mol}{L} = M, \quad [t] = T, \quad (2.32)$$

where we briefly reintroduce the square bracket notation to indicate that we are considering dimensionality;  $M$  is the molar concentration given by the ratio of the amount of substance in moles,  $mol$ , and the liquid volume of the substance in litres,  $L$ , and  $T$  is time dimensionality. Therefore, considering (2.13)

$$\frac{dA}{dt} = -kAB,$$

we require that each term in the equation, including the differential, must be dimensionally consistent such that

$$\left[ \frac{dA}{dt} \right] \equiv [kAB],$$

$$\begin{aligned}
&\Rightarrow \frac{[dA]}{[dt]} = [k][A][B], \\
&\Rightarrow \frac{M}{T} = [k]M^2, \\
&\therefore [k] = \frac{1}{MT}.
\end{aligned} \tag{2.33}$$

Examination of the remaining model ODEs (2.14) and (2.15) provides identical derivation of the dimensionality of the parameter  $k$ . Since the dimensionality of  $k$  does not match that of the dependent and independent variables, we require one or more additional model terms in order to construct parameters with the correct dimensional properties. Consider the dimensionality of  $A_0$ ,

$$[A_0] = M. \tag{2.34}$$

Hence  $A_0$  has the appropriate dimensionality to non-dimensionalise the dependent variables and, when used in conjunction with  $k$ , provides a constant term with the appropriate dimensionality to non-dimensionalise the independent time variable:

$$\left[ \frac{1}{A_0 k} \right] = \frac{1}{[A_0][k]} = \frac{1}{\frac{M}{MT}} = \frac{MT}{M} = T. \tag{2.35}$$

Therefore, we can introduce the following non-dimensional variables,

$$A = A_0 \hat{A}, \quad B = A_0 \hat{B}, \quad C = A_0 \hat{C}, \quad t = \frac{\tau}{A_0 k}, \tag{2.36}$$

where the hat notation represents non-dimensional dependent variables and  $\tau$  represents non-dimensional time. Substituting (2.36) into (2.13) we have,

$$\begin{aligned}
&\frac{d(A_0 \hat{A})}{d(\frac{\tau}{A_0 k})} = -k(A_0 \hat{A})(A_0 \hat{B}), \\
&\Rightarrow k A_0^2 \frac{d\hat{A}}{d\tau} = -k A_0^2 \hat{A} \hat{B}, \\
&\therefore \frac{d\hat{A}}{d\tau} = -\hat{A} \hat{B},
\end{aligned}$$

and by imposing the same substitutions on (2.14) and (2.15) we derive the full non-dimensional basic example model,

$$\frac{d\hat{A}}{d\tau} = -\hat{A} \hat{B}, \tag{2.37}$$



$$\frac{d\hat{B}}{d\tau} = -\hat{A}\hat{B}, \quad (2.38)$$

$$\frac{d\hat{C}}{d\tau} = \hat{A}\hat{B}. \quad (2.39)$$

We must also derive the corresponding non-dimensionalised initial conditions ( $\hat{A}_0$ ,  $\hat{B}_0$  and  $\hat{C}_0$ ), noting that, from (2.36),  $t = 0 \implies \tau = 0$ :

$$\begin{aligned} A &= A_0 \hat{A}, & B &= A_0 \hat{B}, & C &= A_0 \hat{C}, \\ \implies A_0 &= A_0 \hat{A}_0, & \implies B_0 &= A_0 \hat{B}_0, & \implies C_0 &= A_0 \hat{C}_0, \quad (\text{at } \tau = 0) \\ \implies \hat{A}_0 &= 1, & \implies \hat{B}_0 &= \frac{B_0}{A_0}, & \implies \hat{C}_0 &= 0. \end{aligned} \quad (2.40)$$

The non-dimensionalisation process decreases the number of model parameters by one. The parameter  $k$  derived for the original dimensional model is therefore lost and the solution trajectories are now solely dependent on the ratio of the initial concentrations of the dependent variables,  $\hat{B}_0$ . When the initial concentrations of the reactants ( $A_0$  and  $B_0$ ) are equal,  $\hat{B}_0 = \hat{A}_0 = 1$  and hence  $\hat{A}$  and  $\hat{B}$  exhibit the same dynamical response; depleting to zero as the production of  $\hat{C}$  increases to one at steady-state (Figure 2.10A). In contrast, when  $\hat{B}_0 = \frac{1}{2}$  i.e. there is double the

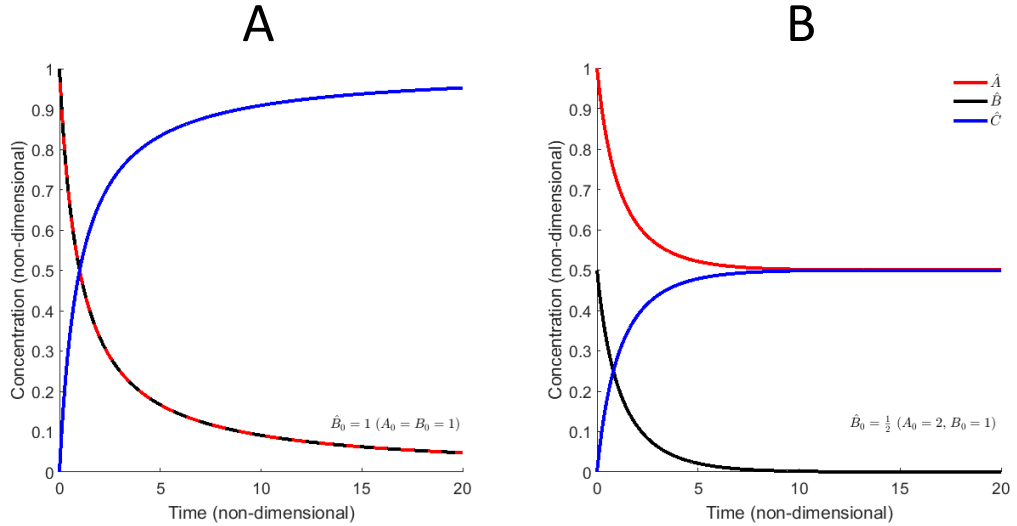


Figure 2.10: Comparison of numerical solutions to the non-dimensional example model. Solution trajectories are plotted for all three non-dimensional dependent variables ( $\hat{A}$ ,  $\hat{B}$  and  $\hat{C}$ ). A) Numerical solutions for the initial condition  $\hat{B}_0 = 1$  ( $A_0 = B_0 = 1$ ). B) Numerical solutions for the initial condition  $\hat{B}_0 = \frac{1}{2}$  ( $A_0 = 2$ ,  $B_0 = 1$ ).

initial concentration of  $A$  as  $B$ , we see distinct dynamical responses for  $\hat{A}$  and  $\hat{B}$  (Figure 2.10B). All of  $\hat{B}$  is depleted in the reaction whereas half of  $\hat{A}$  is depleted since there is half as much of  $\hat{B}$  to interact with. The production of  $\hat{C}$  is therefore restricted and reaches the same steady-state value as the remaining concentration of free  $\hat{A}$ .

## 2.4 Parameter inference

Information regarding model parameter values is very important regardless of whether or not the model is dimensional, since non-dimensional model parameters are invariably comprised of ratios of the original dimensional parameters. Hence, a sound knowledge of the model parameter values or even a general notion of their relative magnitudes can be very valuable in computing reliable model simulations and predictions.

It is, however, highly likely that the relevant reaction rates, translating to model parameters, are shrouded with uncertainty. This is mainly due to the difficulties that arise in recording the appropriate measurements experimentally. Even if one particular reaction is found to be straightforward to characterise, the dimensionality and complexity of mathematical models often generates tens or hundreds of parameter values and so the task of acquiring numerical experimental data for an entire parameter set quickly becomes intractable. There are several approaches to overcoming this obstacle that often involve selecting ‘candidate’ parameter sets, usually at random, and comparing the output generated to a desired response. If a candidate parameter set is able to accurately replicate the desired response then it can be deduced that those particular parameter values are optimal. Parameter inference is the name given to problems of this nature and the techniques designed to solve them can be very useful in analysing model performance.

### 2.4.1 Global optimisation

Establishing reliable parameter sets for making accurate model simulations is difficult without significant investment of resources towards the design and performance of biological experimentation. A number of computational methods exist that look to overcome this limitation, the most reliable of which are global optimisation techniques. Optimisation of a mathematical model involves searching a predefined parameter space for the solution that provides the best fit to experimental observations or an ideal output [Hendrix and Gazdag-Toth, 2010]. That is, the optimal parameter set is located that minimises the error between model outputs and the associated

experimental data. Global optimisation techniques are designed to locate optimal parameter sets globally within often very large parameter spaces, avoiding local minima. One such method, the genetic algorithm (GA), is a particularly powerful global optimisation tool and is exploited regularly in biological model parameter inference [Chen and Chen, 2010; Fernandez et al., 2011]. The GA converges to the global minimum within the allocated parameter space by evolving an initial population of randomly generated solutions over a large number of generations. This process is based on natural selection, giving the best solutions in the population the best chance of creating the next generation of solutions.

Identifying optimal solutions is dependent upon the user-defined fitness function. Fitness functions typically calculate the error between model outputs and experimental data and can be as simple or as intricate as is necessary for obtaining the global optimal solution. The GA function in MATLAB has a wide array of options that can assist the accuracy and speed of a given optimisation run, however many of the default settings are sufficient for the majority of inference problems. As an example, consider our basic model (2.13)-(2.15). The first step is to identify the number of model parameters that will be subject to inference, which in this case is one since our model has only one parameter,  $k$ . Next we select the appropriate population size and number of generations; the MATLAB defaults of 50 and 100 respectively will be sufficient for one model parameter. We then define the size of the parameter space by selecting lower and upper bounds to impose on the parameters; the randomly generated initial population and all subsequent evolved populations examined by the GA can only take values between these bounds. Here we select a lower bound of 0 and an upper bound of 1, noting that we require  $k > 0$  for biological plausibility. We also require data for the model to replicate; for the purposes of this example we will use the synthetic dataset, Concentration 1, in Table 2.1. The synthetic data describes the accumulation of C at eleven time points and hence

Time	0	5	10	15	20	25	30	35	40	45	50
Concentration 1	0.00	0.41	0.58	0.65	0.72	0.75	0.79	0.83	0.84	0.85	0.84
Concentration 2	1.00	0.79	0.71	0.75	0.81	0.85	0.88	0.96	0.93	0.98	0.99

Table 2.1: Synthetic data used to demonstrate parameter inference and model selection. Times and concentrations are given in arbitrary units to provide illustrative observations.

the model output for  $C$  will be matched to this data. Finally, we define the fitness function that will establish the strength of each individual solution. In this case we take the error to be the mean absolute value of the difference between our model

outputs and the synthetic data at the eleven corresponding time points,

$$E = \frac{1}{11} \sum_{i=1}^{11} |x_i - d_i|, \quad (2.41)$$

where  $E$  is the mean error calculated by the fitness function and  $x_i$  and  $d_i$  are the model outputs and data values at each of the relevant time points,  $t_i$ , respectively.

The global minimum mean error computed when the GA terminated was  $E = 0.0095$ , achieved by the corresponding optimal parameter value  $k = 0.1285$ . This considerably low fitness score reveals that our basic model is very capable of replicating the synthetic dataset (Figure 2.11A). The GA converged to this optimal

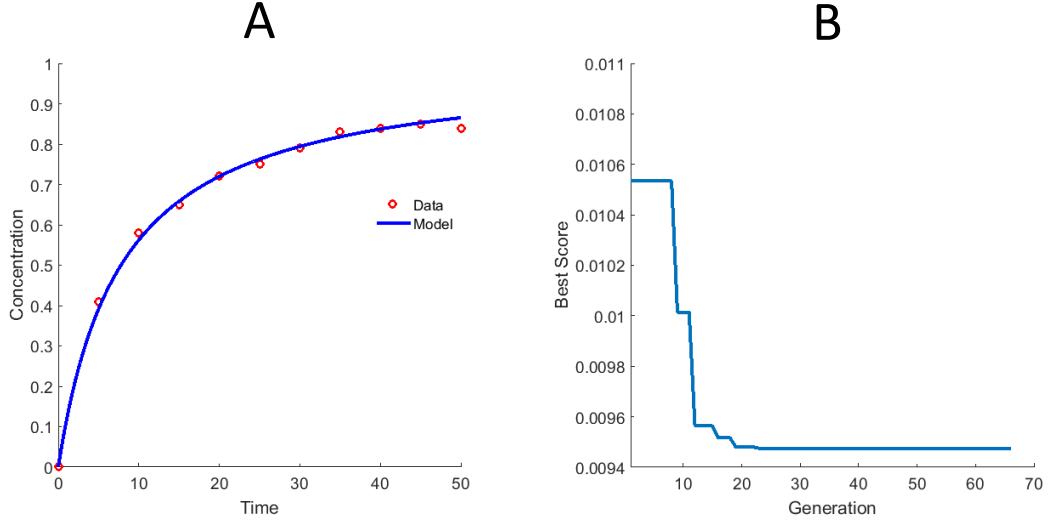


Figure 2.11: Optimal solution and convergence plot for GA global optimisation of the basic model. A) The optimal parameter obtained by the GA,  $k = 0.1285$ , used to simulate the model output for  $C$  against the synthetic dataset. B) The GA converged to the best score (global minimum mean error),  $E = 0.0095$ , after 66 generations.

solution after 66 generations, indicating that the relative tolerance of the fitness function was met (Figure 2.11B). That is, if the average relative change in the best fitness function value over a predefined number of generations is less than or equal to a predefined tolerance then the algorithm will decide that the global minimum has been located and terminates. In this case the predefined values for the generations and tolerance were the MATLAB defaults 50 and  $10^{-6}$  respectively. Note that the GA will terminate based on a number of factors, however, the most common causes are the number of total generations or the relative tolerance, whichever is

reached first. Running the identical optimisation for a second time demonstrates the random nature of the GA. Although a secondary identical GA run obtains almost exactly the same result as before ( $k = 0.1286$ ), the convergence of the algorithm is significantly different (Figure 2.12A). The best score at the first generation is lower

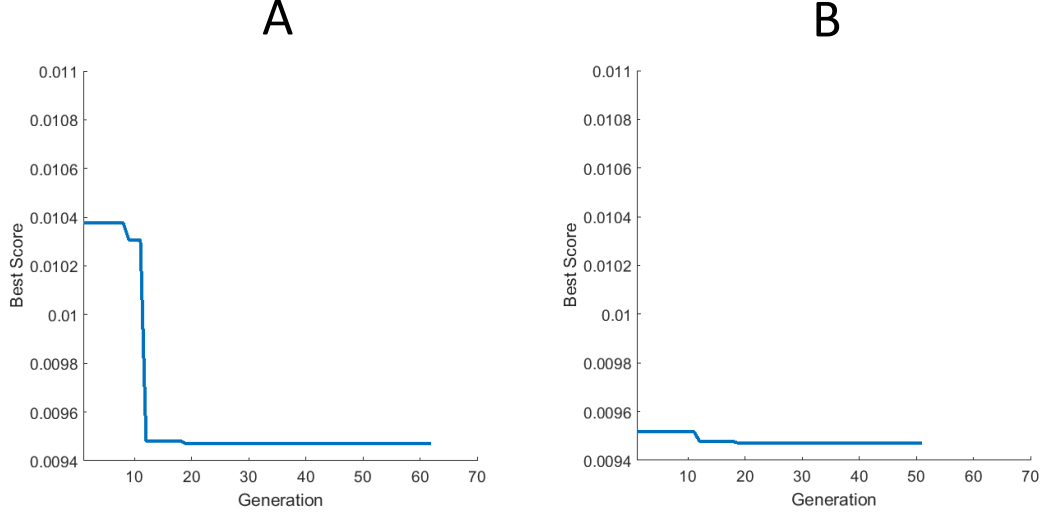


Figure 2.12: Convergence plots for two identical GA optimisation runs. A) A second identical optimisation of the basic model converges to an almost identical optimal solution, but in 62 generations, demonstrating that inherent randomness causes all GA runs to be distinct. B) A third identical optimisation of the basic model, with the exception of parameter normalisation, increases the efficiency of the GA run, converging to an identical optimal solution in 51 generations.

than that of the first run since the initial population is generated randomly and is therefore unlikely to produce the same best scoring solution initially. The process of breeding the solutions to create new generations is also subject to randomness as the population evolves and hence we also see a difference in the number of generations required for the GA to terminate, 62 generations compared to the previous 66. Aside from the optional inputs that can be altered to improve GA performance, we can also normalise the parameter space to enable a more methodical search. Normalising the parameter space between  $-1$  and  $1$  is an effective way of forcing the GA to search for optimal values in the same way for all individual parameters inferred. Imposing all lower bounds of  $-1$  and all upper bounds of  $1$  will force the GA to search  $[-1, 1]$  for all potential solutions. The following calculations of the midpoint,  $m$ , and the mean,  $\mu$ , of the actual parameter set must be made to prevent the GA using the

incorrect normalised values to simulate model outputs:

$$m = \frac{(u - l)}{2},$$

$$\mu = \frac{(l + u)}{2},$$

where  $l$  and  $u$  are the lower and upper bounds imposed on the actual parameter set respectively. The GA can then retrieve the actual values to be used in simulating the model via the following calculation:

$$p_A = mp_N + \mu,$$

where  $p_A$  and  $p_N$  are the actual and normalised parameter sets respectively. Running a third identical GA optimisation of the basic model using this normalisation technique yields a more efficient result than both of the previous runs (Figure 2.12B). The third run obtains the same best score as the second run ( $k = 0.1286$ ) which indicates that the optimal parameter was located in all three runs. This third run does, however, yield a more efficient convergence, requiring only 51 generations to locate the same optimal parameter and global minimum error with respect to the synthetic dataset. Note that normalisation of the parameter space is not greatly advantageous in relatively simple cases such as our basic model as the increased efficiency amounts to relatively small time savings. However, this procedure becomes particularly beneficial when optimising large models with multiple parameters, especially when the bounds imposed on each parameter are varied significantly for reasons regarding the context of the problem. Despite identifying the same optimal solution in each of our three trials, this is not guaranteed in general; multiple ‘optimal’ parameter sets may exist, capable of matching experimental data to a similar degree, depending on the constraints imposed upon the parameter space. Distinct optimal parameter sets can often be indicative of the inference algorithm identifying local minima, in which case several factors may require further consideration such as the sensitivity of the system to changes in parameter values, the size of the parameter space and the quality of the data.

To reiterate, the best optimisation results are achieved for relatively large population sizes and generations compared to the number of parameters subject to inference. However, this also significantly increases the computational workload and hence a reasonable compromise is required for viable development times. Such compromises can still present a very time consuming task and hence GA optimisation is often implemented through parallelisation. Parallel computing allows computa-

tional tasks to be distributed across multiple processors, allowing the workload to be run as several simultaneous jobs. We employ a parallelised MATLAB coding of the GA on a high-performance computing cluster to implement global optimisation of mathematical models in the shortest possible time frames.

### 2.4.2 Approximate Bayesian computation

The analysis of data recorded through biological experimentation naturally comprises the application of statistical methods. Statistics are able to reveal underlying trends and relationships associated with datasets and variables that may not be immediately apparent. Simple statistical measures such as the mean, mode, median, standard deviation and interquartile range (IQR) give a good indication of the spread of data and can also illustrate the extent of any correlations or skewness. Statistical data analysis in biology (biostatistics) can be performed using a host of more advanced methods depending on the information required. Commonly applied methods include the  $\chi^2$  (chi-squared) test which calculates the deviation of a dataset from the expected values of the null hypothesis, that is, assuming that there is no effect of a given factor on the dataset, how much deviation from ‘no effect’ is actually caused by that factor, and the Spearman’s rank correlation coefficient which generates a number between 1 and  $-1$  that represents the strength of the relationship between two variables when described by a monotonic function. That said, these methods can only be applied directly to data and do not offer any means of measuring correlations that might exist between experimental observations and model outputs. We are, however, able to exploit the statistical certainty of the relationship between experiment and computation via parameter inference techniques based on probability. Such methods are also capable of model selection, the probabilistic discrimination of two or more models of the same system, which can identify the model most likely to have produced the associated experimental data.

Probability theory is fundamental to statistical analysis in providing rigorous means of determining the likelihood of certain events occurring. The probabilistic relationship between two events can be depicted in the form of a Venn diagram (Figure 2.13); the probability of event A,  $P(A)$ , is depicted by the left hand circle, the probability of event B,  $P(B)$ , is depicted by the right hand circle and the probability of both events occurring,  $P(A \cap B)$ , is depicted by the central area of intersection between the two circles. Two events that are mutually exclusive, such as a coin toss, cannot both occur and hence there is no intersection in these cases;  $P(A \cap B) = 0$ . The area of the outer rectangle containing the intersecting circles represents the entire probability and is therefore equal to 1. For mutually non-exclusive events

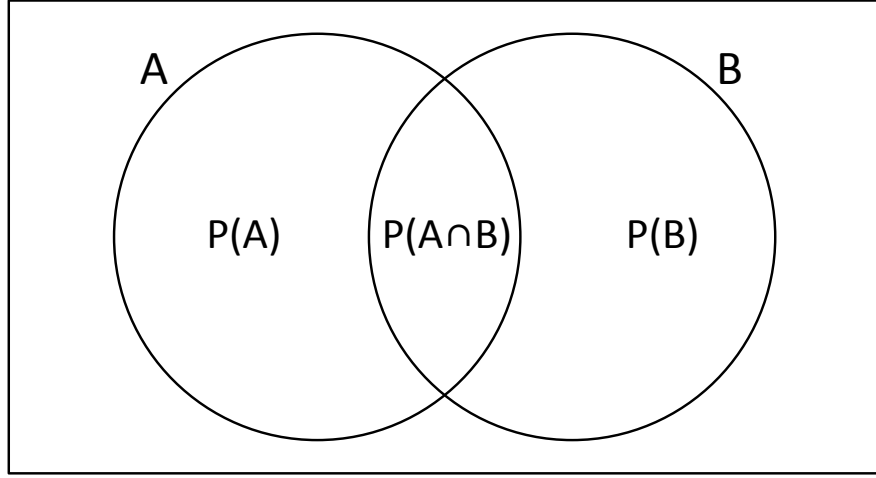


Figure 2.13: Venn diagram depicting the probability of two events, A and B, by virtue of the areas of the intersecting circles. The areas of the two circles are equal purely for illustrative purposes.  $P(A)$  and  $P(B)$  denote the probabilities of event A and B occurring respectively;  $P(A \cap B)$  denotes the probability of both events occurring and corresponds the area of intersection between the two circles.

which have, or are at least assumed to have, some shared dependency, we can calculate the probability of one event occurring given the prior occurrence of another event by virtue of conditional probability theory:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad (2.42)$$

where  $P(B) > 0$  and  $P(A|B)$  denotes the conditional probability of event A given event B, which is equal to the ratio of the probability of the intersection between the two events and the probability of event A. That is, if  $P(A \cap B)$  is relatively large, there will intuitively be a high probability of A occurring in the event that B has occurred however, if  $P(A \cap B)$  is small then it becomes more unlikely that A will occur given B has occurred. Conditional probability is axiomatic in the application of probability theory since it provides a platform for accounting for any preconceived relationships or dependency that might exist between events of interest. Consider the probability of B given A:

$$P(B|A) = \frac{P(B \cap A)}{P(A)}. \quad (2.43)$$

The intersection of A and B,  $P(A \cap B)$ , and the intersection of B and A,  $P(B \cap A)$ , are the same (Figure 2.13) and hence,



$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \quad (2.44)$$

which we can rearrange to determine the intersection of A and B in terms of conditional probability:

$$P(A \cap B) = P(B|A)P(A). \quad (2.45)$$

Substituting (2.45) into (2.42) gives the full conditional probabilistic relationship between two mutually non-exclusive events known as Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \quad (2.46)$$

where  $P(B) > 0$ . This theorem forms the basis of Bayesian inference, a potent class of statistical inference techniques that facilitates the parameterisation of mathematical models through the imposition of prior knowledge on the problem, usually in the form of experimental data, and is therefore well suited to mathematical modelling investigations in biological systems research. Bayesian inference enables us to calculate the probability that a particular parameter set,  $\theta$ , gives rise to a given experimental dataset,  $D$ , by virtue of Bayes' theorem:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}, \quad (2.47)$$

where  $P(\theta|D)$  and  $P(\theta)$  are referred to as the posterior distribution and prior distribution respectively. Probability distributions such as the posterior and prior in Bayes' theorem are functions that describe the probability that the subject variable will take a possible value. This applies to random variables and the appropriate probability distributions take on different forms depending on whether the variable is discretely random or continuously random [Ross, 2010]. In this case, we are only interested in varying  $\theta$  with respect to our observed data,  $D$ , which is fixed. Hence,  $P(D)$  is constant and has no meaningful influence on the overall inference of the posterior distribution, resulting in the following proportionality:

$$P(\theta|D) \propto L(\theta|D)P(\theta), \quad (2.48)$$

where  $L(\theta|D)$  is referred to as the likelihood function. The elucidation of the posterior distribution is dependent on the likelihood function and the prior distribution

and forms a general result of probability theory:

$$\text{posterior distribution} \propto \text{likelihood function} \times \text{prior distribution}. \quad (2.49)$$

The likelihood is a function of the model parameters that determines the probability that a particular parameter set is responsible for producing the observed data, whereas the prior distribution explicitly prescribes the biological knowledge regarding the model parameters before the relevant experiments were carried out [Liepe et al., 2014]. Evaluation of the posterior distribution is typically performed through approximate Bayesian computation (ABC). The likelihood function is comparable to the fitness functions used in parameter optimisation strategies such as the GA however, it can only be expressed explicitly for appropriately simple models and hence ABC expands the scope of model analysis by bypassing the need to evaluate such functions. Instead, bespoke computational tools such as ABC-SysBio can be employed to obtain the desired result. ABC-SysBio is a Python software package that is designed specifically for parameter inference and model selection in biological systems research [Liepe et al., 2014]. The program enables ABC inference of mathematical models via sequential Monte Carlo (SMC) approaches [Liepe et al., 2014; Stumpf, 2014; Smith and Grohn, 2015]. Monte Carlo approaches to computational simulations involve generating random candidate solutions, testing their fitness against a desired output and repeating until a viable solution can be identified. In this way, vast numbers of randomly selected parameter sets can be examined in building an accurate approximation to the posterior distribution. The methods were developed by Stanislaw Ulam and John von Neumann in conjunction with top secret work relating to the development of nuclear weapons. As such, a code name was required and the term Monte Carlo was adopted after the famous Monte Carlo casino in Monaco, due to the connection between randomness and games of chance [Harrison, 2010].

The ABC-SMC approach proceeds in the following manner: the first ‘population’ of accepted solutions or ‘particles’ is generated randomly based on the prior distributions imposed on the model parameters. Each particle gives rise to a simulated dataset,  $D^*$ , which is compared to the fixed experimental dataset,  $D$ , by an appropriate distance function and its fitness is scored accordingly. Acceptance of a particle is dependent on a decreasing sequence of error thresholds,  $\epsilon$ , set to correspond with each population. That is,

$$d(D^*, D) < \epsilon_i, \quad (2.50)$$

where  $\epsilon_1 > \epsilon_2 > \dots > \epsilon_n$  and  $d$  is the distance function. Each subsequent population is obtained by perturbing particles from the previous population in accordance with a predetermined perturbation kernel, proceeding until the model is unable to produce particles of sufficient fitness to satisfy the immediate threshold.

An array of model-specific criteria are required to allow the ABC-SysBio package to run efficiently: the sequence of decreasing error thresholds,  $\epsilon$ , must be provided whereby only the particles capable of providing error less than that of the threshold will be accepted by the algorithm. Each  $\epsilon$  must be satisfied in succession until the particles are unable to satisfy the next threshold. Satisfaction of an individual threshold is dependent on the number of particles accepted; the number of acceptable particles required before passage to the next threshold must also be predetermined. The larger the number of particles, the higher probability of significant inference and the longer the duration of algorithm to reach convergence. Each individual parameter subject to inference requires a prior probability distribution in order to establish the parameter space within which to locate acceptable particles. Sequences of numerical values representing the relevant experimental data and the corresponding time points must also be provided; the number of data points and time points must be equal. Time course data is currently the only supported data format. One or more distinct datasets can be supplied and can be fitted to any individual model variable or combination of variables. Convergence of the algorithm is dependent on all of the aforementioned factors and hence it may require several trials to establish the appropriate performance criteria. It is advised that strong results are repeated multiple times due to the random nature of the Monte Carlo simulations that drive the algorithm. Note that all models submitted to the ABC-SysBio package must be written in Systems Biology Markup Language (SBML), a systems biology programming language based on Extensible Markup Language (XML).

Implementing ABC-SysBio for one model only initiates parameter inference, whereas submitting more than one model will initiate model selection. Note that implementing model selection in ABC-SysBio also conducts parameter inference on each model subject to selection. At each satisfied error threshold, the algorithm produces a host of output results including line graph simulations of ten example particles from the accepted population of particles, histograms of the posterior distributions relating to each model parameter and, in the case of model selection, histograms detailing model probabilities and the corresponding population number, error threshold and acceptance rate. In order to demonstrate parameter inference in ABC-SysBio we run the algorithm for the same basic model (2.13)-(2.15) that we

optimised using the GA. The  $\epsilon$  thresholds were selected knowing that the optimal GA solution was found to provide a minimal error of 0.0095, that is, the final  $\epsilon$  was selected as 0.007 however, this information would not typically be available and hence an informed estimate of  $\epsilon$  thresholds is required. The number of accepted particles required to satisfy each threshold was selected as 1000 to achieve convergence within a viable time frame. The prior distribution imposed on the model parameter  $k$  was selected as a uniform distribution on the interval  $[0, 1]$ . The final population of 1000 accepted particles satisfied  $\epsilon = 0.01$  however, the following threshold,  $\epsilon = 0.009$ , was not satisfied which indicates that ABC-SysBio aligns with the minimum error achieved by the GA. The subset of ten particles plotted at the last satisfied threshold all show a close match to the synthetic data (Figure 2.14A); the threshold was satisfied after sampling 5354 particles, giving an acceptance rate

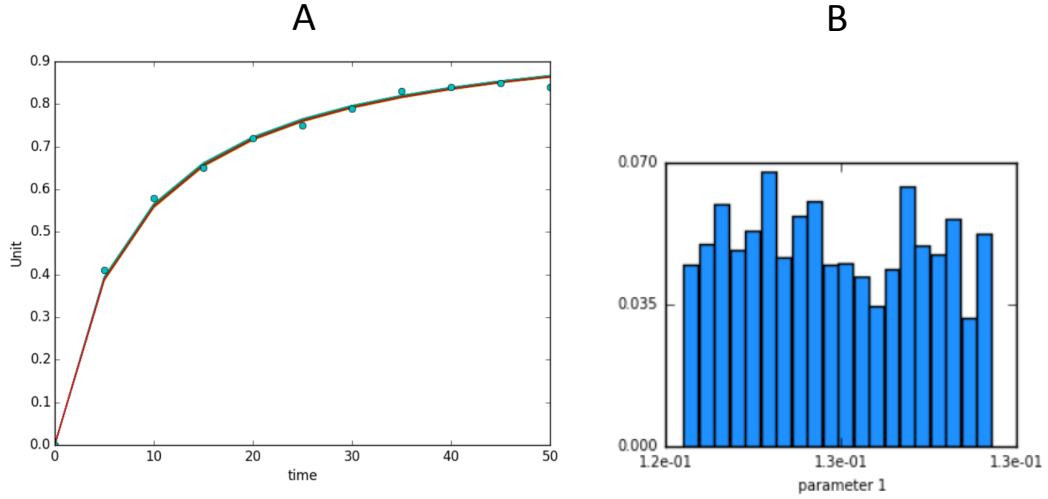


Figure 2.14: ABC-SysBio parameter inference results. A) A subset of ten solutions from the final population of accepted solutions all provide a close match to our synthetic data. B) The inferred probability distribution for the model parameter,  $k$ , reveals a high probability that the optimal value is located in the interval  $[0.12, 0.13]$ .

of approximately 0.19 i.e. less than 20% of sampled particles were accepted. The inferred probability distribution on the model parameter  $k$  indicates that the probability is entirely distributed across the interval  $[0.12, 0.13]$ , aligning with the optimal value of 0.1285 achieved by the GA (Figure 2.14B). The distance function deployed to determine particle fitness is customisable and was selected here to be the mean absolute error to emulate the GA fitness function (2.41).

In order to demonstrate model selection in ABC-SysBio, we require two or more distinct models of the same biological system [Liepe et al., 2014]. We there-

fore submit the aforementioned mechanistic and black box models of enzyme kinetics which will also assist in establishing further evidence of the advantages and disadvantages associated with each modelling strategy. The criteria selected for the previous demonstration of parameter inference, as well as the mean absolute error distance function, are retained for this model selection with the following exceptions: the synthetic dataset is retained since the time course evolution of product is expected to be similar to that of the basic model however, we also introduce a second dataset to fit the time course evolution of enzyme (Table 2.1, Concentration 2). Enzyme concentration is depleted through complex formation with the substrate, but will replenish as the substrate is converted to product and hence the data values are chosen to reflect this. Since we are inferring against the fit to two distinct datasets, we also adapt the  $\epsilon$  sequence to account for greater potential error. The model probabilities are plotted as histograms for each satisfied error threshold, in this case, producing twelve subplots (Figure 2.15A). Both models have approximately equal

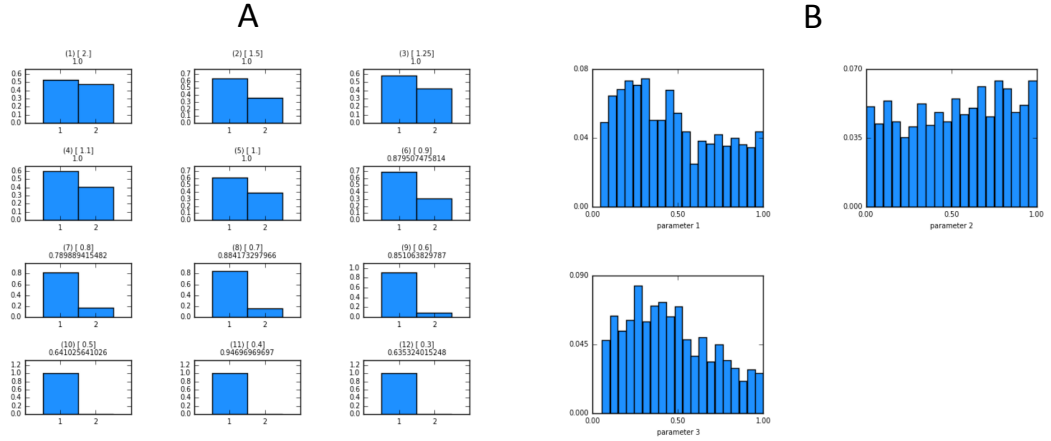


Figure 2.15: ABC-SysBio model selection results. A) The mechanistic model achieves a probabilistic certainty of producing our two synthetic datasets with a minimum error of 0.3. B) The inferred probability distributions for the three mechanistic model parameters reveal that parameters 1 and 3 take optimal values in the interval  $[0, 0.5]$  and parameter 2 is optimal in the interval  $[0.5, 1]$ .

probability of producing the data after the first threshold,  $\epsilon = 2.0$ , however, by the tenth threshold,  $\epsilon = 0.5$ , the mechanistic model (model 1) has achieved a probabilistic certainty of producing the data and an eventual minimum error of  $\epsilon = 0.3$ . This is unsurprising when considering that the black box model overlooks the intermediate formation of the enzyme-substrate complex and, in turn, the dissociation of enzyme that enables its concentration to replenish as the system reaches steady state. The inference of the three mechanistic model parameters is not as clear as

the basic model example given the increased difficulty of the twofold data fitting task (Figure 2.15B). The probabilities are distributed across the entire interval prescribed as the prior distributions however, general trends can still be identified such as greater probabilities that the optimal values are located in the interval  $[0, 0.5]$  for parameters 1 and 3 ( $k_1$  and  $k_2$ ) and in the interval  $[0.5, 1]$  for parameter 2 ( $k_{-1}$ ). These values suggest that the system requires lower magnitude (slower) reaction rates in order to mimic the synthetic data on the given timescale. Overall, this demonstration has revealed that the mechanistic model not only achieves a greater probability of capturing the desired dynamics, but that it is a statistical certainty and illustrates the importance of accounting for the maximum amount of biological detail possible in formulating mathematical models.

Bayesian inference provides an effective model validation method whilst eliminating problems regarding over-fitting and uncertainty that have the potential to compromise optimisation strategies [Liepe et al., 2014]. Although the optimal parameter set is generally considered to be the most desirable outcome of a given inference problem, optimisation techniques with this sole focus are susceptible to placing too much trust in the available data. That is, if the experimental data used to optimise a model of a biological system contains any element of noise, which is invariably true, then a good optimisation algorithm will work efficiently to fit the model to that data, as required, and inadvertently locate noisy parameter sets. Optimisation is also prone to locating local minima, creating uncertainty as to whether the identified solution is truly optimal. Furthermore, even trusted optimal solutions could potentially be one of many such solutions that provide equally minimal error and hence a great deal of consideration is required with respect to the robustness and biological plausibility of solutions, as is the case for interpreting any inference result. ABC methods avoid these issues by virtue of their probabilistic nature; the approximate posterior distributions on the inferred parameters are measures of certainty in themselves and offer a probable subset of the parameter space, rather than a definitive output to be accepted or rejected. That said, both the definition of the relevant parameter space and the subsequent exploration of it remains the nucleus of inference strategies and is arguably confronted with greater sophistication through optimisation, in comparison with ABC-SMC approaches. As a global optimisation strategy, the GA is designed specifically to avoid convergence towards local minima, with issues of uncertainty and robustness combated by implementing multiple runs under identical performance criteria; another general feature of all parameter inference methods. Both inference strategies outlined in this section possess several potential pitfalls which in many cases are only effectively overcome through

astute implementation and interpretation. Overall, we have established a reliable, two-pronged tool kit with which to parameterise mathematical models and, in turn, elucidate biological systems.

## Chapter 3

# Mechanistic Modelling of a Rewritable Recombinase Addressable Data Module

### 3.1 Scientific background

#### 3.1.1 DNA recombination

Genetic switches form the basis of engineering cellular memory [Bonnet et al., 2012]. Transcriptional memory devices have demonstrated effective performance across multiple cellular environments [Gardner et al., 2000; Kramer et al., 2004b] and are highly orthogonal with regards to assembling multiplexed systems [Stanton et al., 2014], however regulating gene expression in this manner also has limitations. These systems are volatile, having to continuously consume resources, in this case for the production and degradation of repressor, to maintain states [Ajo-Franklin et al., 2007]. Difficulties also arise when integrating devices into a variety of organisms given that gene regulation networks vary greatly between distinct cellular environments. Furthermore, the highly inducible and stable switching that these devices demand can be compromised by spontaneous switching events caused by the inherent stochasticity of gene regulation [Bonnet et al., 2012]. As a result, research into cellular memory has become increasingly focused towards site-specific recombinases (SSRs), capable of precise DNA manipulation both *in vitro* and *in vivo* [Grindley et al., 2006]. This process is known as DNA recombination and facilitates inducible gene expression that is programmed into the cellular DNA.

DNA-based systems are favourable due to the fact that they exploit a natural data storage material and have the added advantage of eliminating cell specificity



requirements. SSRs are classified as belonging to two groups, the tyrosine recombinases and the serine recombinases. The former have been shown to provide effective genetic switch mechanisms [Buchholz et al., 1996; Kilby et al., 1993; Rossant and Nagy, 1995] however, their functionality is often dependent upon cell-specific cofactors as in the case of  $\lambda$  integrase [Landy, 2015]. This is problematic in a similar vein to that of transcriptional systems with regards to the deployment of modules across multiple organisms. Tyrosine recombinase systems that are not dependent on host cofactors are bidirectional and are therefore incapable of highly efficient switching since there can be no guarantee that an induced transition to a desired DNA state would be effectively maintained without unwanted transitioning back to the original state [van Duyne, 2001]. In contrast, the serine recombinases do not require such cofactors and have been used effectively to perform highly efficient unidirectional gene assembly and modification [Colloms et al., 2014]. This has led to the construction of a rewritable RAD module exhibiting passive information storage within a chromosome [Bonnet et al., 2012]. Switching the RAD module ‘on’ requires only the presence of integrase whereas the ‘off’ switch requires integrase in conjunction with a recombination directionality factor (RDF), also referred to as excisionase. The integration and excision events can take on several guises depending on the initial composition of the genetic sequence and the relevant attachment sites. Specific recombination arrangements are directly influential upon the operational characteristics of a genetic toggle switch, such as the RAD module, and higher-order circuitry that utilises multiplexed switches.

Three distinct DNA recombination mechanisms are known to mediate three distinct recombination events, referred to as inversion, insertion and deletion. The first mechanism exploits the inversion event (Figure 3.1A). In this case, antiparallel attB and attP sites located on the same DNA sequence are subject to binding by dimeric integrase which causes double stranded breaks in each. That is, integrase alone is sufficient to mediate a primary inversion event. Exposed ends in the genetic sequence then bind the opposite ends of the intermediate fragment, resulting in an inverted section of DNA flanked by newly formed composite attachment sites termed attL and attR. The binding of RDF molecules to the attL and attR DNA:integrase synaptic complexes facilitates a successive inversion event. The attL and attR sites are dismantled, allowing for the exposed ends of the intermediate fragment to adopt their original orientation and hence re-establish the attB and attP sites.

The second mechanism exploits the deletion and insertion events between attachment sites on the same DNA sequence (Figure 3.1B). In this case, the attachment sites are identical to that of Figure 3.1A with the exception that the orientation

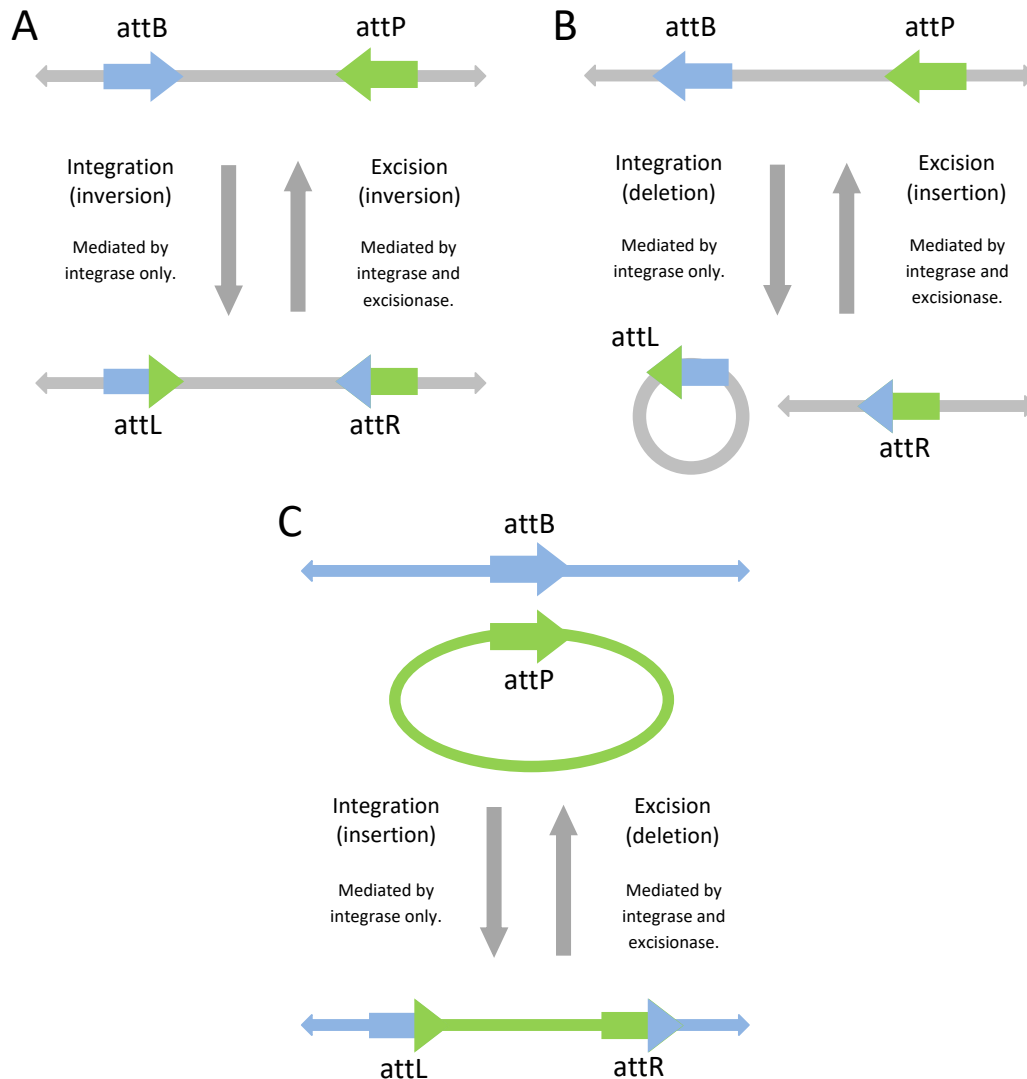


Figure 3.1: Schematic diagram of DNA recombination mechanisms. A) Genetic material flanked by attB and attP is inverted through integration to form attL and attR. Excision restores attB and attP via a secondary inversion event. B) An alternative orientation of attB and attP results in deletion of the intermediate genetic sequence which is inserted in the presence of RDF (excisionase). C) The phage genome attachment site, attP, is integrated into the host chromosome attachment site, attB. Excision restores attB and attP, removing the integrated phage genome from the host chromosome. In all cases, integration gives rise to attL and attR, each formed of half of attB and attP.

of attB and attP is subtly different i.e. they are parallel as opposed to antiparallel, akin to attL and attR in Figure 3.1C. Binding of dimeric integrase to attachment

sites adopting this alternative orientation causes the same double stranded breaks however, the exposed ends of the intermediate nucleotide sequence are bound together to form a small loop of DNA that is deleted from the original sequence. The exposed ends of the original sequence are also bound together to form a single attR site. RDF binding re-inserts the deleted loop of DNA and reforms attB and attP. Hence, this alternative orientation of attachment sites does not facilitate the same inversion of the intermediate genetic fragment depicted in Figure 3.1A, but instead deletes the fragment entirely with the potential of insertion via an appropriate RDF.

The third mechanism also exploits insertion and deletion. In this case, recombination occurs between DNA attachment sites on a bacterial host chromosome and a bacteriophage (Figure 3.1C). Dimeric integrase bound to a host chromosome attB site and a bacteriophage attP site causes a double stranded break in each. The exposed ends of the phage DNA fragment bind to those of the host, thus inserting the fragment into the host chromosome which is flanked by the newly formed composite attL and attR sites. Following insertion, binding of RDF molecules to the attL and attR DNA:integrase synaptic complexes facilitates the deletion event. Here the attL and attR sites are dismantled, thus deleting the genetic insert and allowing the reformation of the attB and attP sites.

Inversion events giving rise to attL, attR sites and attB, attP sites are generally referred to as integration and excision respectively since, in the main, integrase alone mediates the former and a combination of integrase and RDF is necessary to mediate the latter. In contrast, insertion and deletion events can both potentially be mediated by integrase alone or integrase and RDF, and are therefore associated specifically with their aforementioned DNA recombination outcomes. We investigate inversion- and deletion-based DNA recombination systems since the action of these mechanisms is localised to a single DNA strand which offers greater simplicity compared to the manipulation across disparate genetic sources associated with insertion. We refer to recombination events as integration and excision according to the attachment sites and SSRs involved and to the presence of attB, attP sites and attL, attR sites in the system as the BP and LR states respectively. The concentrations of integrase and RDF in the system dictate the efficiency of the switching between these distinct DNA states. The characterisation of this switching efficiency is crucial to the deployment of the genetic switch as a component of higher-level systems. By analysing the criteria required to mediate both highly efficient integration and excision *in silico*, optimal system inputs can be identified.

### 3.1.2 Existing recombinase-based systems

Many of the earliest publications regarding both tyrosine and serine integrases were focused on elucidating their structural and functional properties through comprehensive experimentation [Thorpe and Smith, 1998; Thorpe et al., 2000; Smith and Thorpe, 2002; Ghosh et al., 2003; Groth and Calos, 2004]. In light of these observations and the continued publication of increasingly detailed studies regarding the nuances underlying DNA recombination, such as the function of the synaptic complex [Rowley and Smith, 2008; McEwan et al., 2009, 2011; Bai et al., 2011; Olorunniji et al., 2012], a slew of useful applications were identified that prompted the design and implementation of many novel systems. Serine integrases have become highly prevalent as integration vectors, emerging as the preferred method of transferring DNA into other organisms such as streptomycetes [Zhang et al., 2013]. These SSRs are capable of mediating integration of entire antibiotic biosynthetic clusters into target genomes [Baltz, 2011, 2012] and are also highly compatible when interconnected within the same organism [Gregory et al., 2003].

Other applications of the serine integrases are focused on therapeutics such as engineered bacteria that are programmed to invade cancer cells [Anderson et al., 2006] and the production of two essential human blood-clotting proteins known as factor XII and factor IX in mice to treat haemophilia A and B respectively [Chavez et al., 2010; Olivares et al., 2001]. Additional biomedical applications include transgenic cattle capable of expressing milk containing the human  $\beta$ -defensin-3 antimicrobial peptide which naturally protects the surfaces of human organs and blood vessels from bacterial colonisation [Yu et al., 2013], inducible production of pluripotent stem cells from human amniotic cells and embryonic cells in mice [Ye et al., 2010] and engineering specific skin cells and partially specialised stem cells for gene therapy of skin disorders [Ortiz-Urda et al., 2003, 2002].

Sophisticated methods concerning the assembly and optimisation of complex large-scale synthetic systems are ideally suited to site-specific DNA recombination [Xu et al., 2013]. DNA assembly via serine integrases enables the construction of highly modular pathways that can be adapted without the need for repeated cloning [Zhang et al., 2008, 2011] and can also facilitate rapid metabolic pathway assembly [Colloms et al., 2014]. Genome engineering is also benefiting from the application of serine integrases;  $\phi$ C31 has been utilised for the specific modification of genomes in mice [Tasic et al., 2011], silkworm embryos [Yonemura et al., 2013, 2012] and zebrafish [Hu et al., 2011; Lister, 2010], thus elucidating gene function in model organisms, and also for the deletion of genetic markers in plants such as *Arabidopsis* (rockcress), wheat and barley [Thomson et al., 2010; Kempe et al., 2010;

Kapusi et al., 2012], thus generating stable progeny void of undesired DNA.

The potential of recombinase-based circuitry to provide transistor-like behaviour in synthetic biological systems has naturally opened up numerous applications relating to genetic data storage and biocomputing [Baumgardner et al., 2009; Ham et al., 2008]. Digital data storage in particular has attracted much attention, resulting in several validated memory devices such as the aforementioned event counter that utilises tyrosine integrase-based recombination to count and record sequential pulses of inducer; the purely transcriptional circuit tested in parallel was only able to count induction events [Friedland et al., 2009]. Of course, transcriptional elements are necessary for realising desired outputs however, these results indicate that their control must be mediated via DNA recombination in order to achieve memory of external stimuli. Memory modules can support efficient inducible DNA inversion in alignment with mathematical modelling simulations [Bonnet et al., 2012], store over 1 B of digital information via layered arrangements of attachment sites specific to multiple recombinases [Yang et al., 2014] and represent integral components in the construction of a biological microprocessor [Moe-Behrens, 2013]. Consequently, the level of input-output complexity that can be realised is theoretically unbounded due to the scaling up of systems through numerous pairs of attachment sites corresponding to distinct integrases, therefore providing the platform for engineering the full range of Boolean logic operations in response to induction of independent serine integrases [Bonnet et al., 2013; Siuti et al., 2013]. At the current rate of expansion, it is thought that worldwide data will require in the region of  $4 \times 10^{10}$  (forty-trillion) GB of digital storage by 2020 and, although cloud computing is hoped to address this demand, approximately 90 grams of DNA would be sufficient to store such an amount [O’Driscoll and Sleator, 2013].

Of the diverse range of recombinase-based systems, only a small fraction have been published in conjunction with mathematical modelling investigations that reveal, for example, specific reaction rates that are currently intractable experimentally, or expected dynamical behaviour via qualitative or even quantitative *in silico* simulations as a reference point for *in vivo* circuit assembly [Ringrose et al., 1998; Bonnet et al., 2012; Friedland et al., 2009]. Of these extant models, even fewer are related specifically to serine integrase recombination interactions which may be surprising considering the relative wealth of publications detailing the structural and functional properties of recombinases that could inform mechanistic model construction. It is worth reiterating that the repressilator and toggle switch were both designed and characterised in alignment with albeit simple mathematical models. The deployment of synthetic biological devices is unlikely to ever be entirely predictable

purely by virtue of experimental observations and thus there exists a definite need for extensive modelling approaches to serine integrase reactions in order to deliver and enhance the aforementioned applications.

### 3.1.3 Existing recombinase-based models

As cellular memory emerges as a defining element of higher-level synthetic biological systems, the characterisation of the requisite parts will command significant attention. Hence, predictive analysis of recombinase-based genetic switches is necessary to provide engineers with reliable operational profiles. Achieving this goal is possible through the inevitable progression in the efficacy of experimental procedures due to technological advancements. However, mathematical modelling approaches have the potential to provide insights that may never be physically possible in the laboratory. Wider acceptance of the merits of mathematical models in biology and increased efforts to expand collaborative experimental and computational research is therefore central to synthetic biological circuit design.

The earliest proposed model of DNA recombination in the literature provides kinetic analysis of two distinct recombinases known as FLP (flippase) and Cre [Ringrose et al., 1998]. The model captures a simplistic overview of DNA deletion via a series of reversible reactions corresponding to monomeric recombinase binding to DNA attachment sites, synaptic complex formation, recombination and dissociation. The universal reversibility of the reactions modelled aligns with FLP and Cre being tyrosine recombinases that exhibit bidirectional recombination and hence the model also captures the insertion of the circular DNA product back into the genetic sequence. Two pairs of reactions corresponding to two distinct aspects of synapse formation were unable to be determined experimentally, resulting in four unknown model parameters; the remaining four model parameters were established via experimentation. The model was optimised to infer the four unknown parameters through fitness function minimisation. As a result, a number of dynamical properties were validated including that Cre has a higher binding affinity than FLP and thus the synaptic complex is more stable for Cre which was thought to explain the 100% deletion efficiency of FLP compared to the maximum 75% excision efficiency of Cre. Furthermore, insertion was shown to be inefficient given the bidirectionality of tyrosine-mediated recombination that results in unwanted ‘re-deletion’ and, although it may seem intuitive that insertion efficiency would benefit from increased DNA binding affinity and rate of synapsis, such conditions are in fact detrimental to insertion efficiency since they favour re-deletion [Ringrose et al., 1998]. Therefore, this modelling investigation succeeded in highlighting both the optimal model

parameter set and the operational faults that render the mechanism unsuitable for use as a cellular memory unit.

A validated model of serine DNA recombination did not arrive until fourteen years later [Bonnet et al., 2012]. This serine model adopts a simplistic, black box approach akin to that of [Ringrose et al., 1998] however, in this case, the specific focus is DNA inversion as opposed to deletion and insertion. The model captures *in vivo* DNA recombination reactions relating to the serine integrase gp35 and the RDF gp47 from the bacteriophage Bxb1 and is referred to as a rewritable RAD module. In contrast to their tyrosine cousins, serine recombinases mediate unidirectional recombination which makes them conducive to inducible regulation of gene expression [Bonnet et al., 2013].

The model accounts for the dynamical behaviour summarised in Fig. 3.2: integrase and RDF are both expressed in monomeric form in solution, but integrase

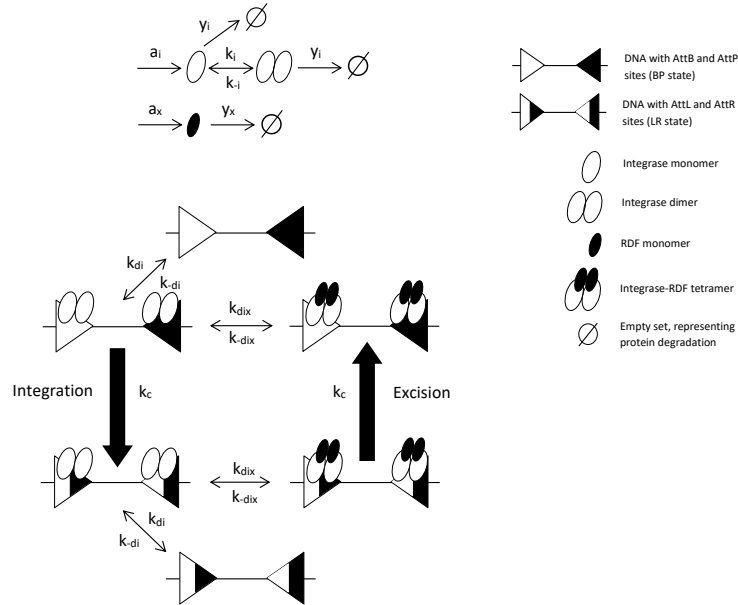


Figure 3.2: The RAD module DNA recombination reaction network taken from Bonnet et al. [2012].

alone undergoes dimerisation in solution. That is, pairs of integrase monomers bind together in solution to form dimeric complexes. Integrase dimers are then able to bind specifically to attB and attP sites on the free DNA substrate. One integrase dimer bound to each attachment site is necessary and sufficient to mediate the primary inversion event, referred to as integration since it involves integrase only. This causes the double stranded break in the DNA and the subsequent re-

ligation of opposing ends of the intermediate genetic fragment that gives rise to the composite attL and attR sites. RDF monomers only bind to integrase dimers already bound to DNA, forming a synaptic tetramer. The binding of two RDF monomers to each integrase:DNA synaptic complex is necessary and sufficient to mediate the secondary inversion event, referred to as excision since both integrase and RDF are involved. RDF is also able to bind integrase:DNA complexes in the BP state, thus inhibiting integration; there is no RDF binding to integrase in solution [Bonnet et al., 2012]. The BP and LR DNA states are tagged with GFP and RFP respectively to provide clear readout of the recombination efficiency of the system. The model describes *in vivo* DNA recombination and, as such, the expression and degradation of recombinase proteins represent a key dynamical element; *in vitro* studies such as [Ringrose et al., 1998] are void of environmental pressures and hence concentrations of recombinases are synthesised experimentally. In all, the model is comprised of nine distinct variables and eight parameters representing the integration-excision DNA inversion mechanism.

The model was used to identify the operational properties of the RAD module required for delivering digital information storage. The key feature of the system is the efficiency of switching between DNA states. This is ascertained by establishing the concentration every molecular entity in the DNA state of interest and computing the evolution of the summed total over time. The sum of the relevant concentrations is referred to as the total register of the system and is calculated *in silico* by summing each ODE corresponding to molecular entities in the same DNA state. Experimentally, the total register of the system is measured directly as the intensity of GFP or RFP output, with increased fluorescence signifying increased recombination efficiency. Presuming that the module initially adopts the BP state, the ‘set’ operation constitutes an efficient ‘on’ switch transitioning the system to the LR state via induced integrase and an absence of excisionase; the ‘reset’ operation therefore constitutes an efficient ‘off’ switch that reproduces the initial BP state via induced integrase and RDF.

The ability of the system to demonstrate robust ‘hold’ states, whereby the most recent set or reset operation is maintained in the absence of inducer, is another important criterion with regards to temporal control of the module. Switching efficiency is dependent on the concentrations of integrase and RDF in the system however, these quantities are difficult to determine numerically. As a result, the relative ratios of recombinase expression and degradation rates were examined with respect to the percentage switching efficiency they produced. A vast array of ratios were tested to provide expected dynamical responses with which to direct exper-



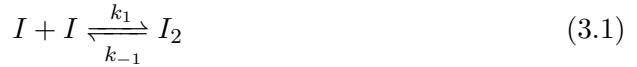
imentation. That is, no experimental data was used to inform the selection of individual parameter values or to provide specific outputs for model validation purposes, and thus the model was suitably non-dimensionalised to offer operational conditions to be emulated experimentally. All model parameters corresponding to protein:DNA interactions were set equal to 1 with the basal and induced ratios of recombinase expression and degradation rates set at 0.1 and 10 respectively. RAD module operations were initially investigated in separation; set operations were isolated with approximately 95% switching efficiency and a robust hold state in the absence of RDF, which is intuitive given that the set function encompasses the integration reaction that is solely mediated by integrase. Isolated investigation of the reset operation revealed reversibility *in vivo* despite experimental evidence of approximately 100% switching efficiency regarding Bxb1 integrase and excisionase *in vitro* [Ghosh et al., 2006]. This highlights the influence of noisy cellular conditions on recombinase expression and degradation capable of causing re-integration of the reset inverted genetic sequence that would be observed as reversibility and, ultimately, system failure.

A full set-reset cycle was also attempted however, it was observed that the majority of set and reset functions that exhibited high efficiency in isolation were unable to give rise to sufficiently robust full cycles required of a digital storage module. For example, the high concentrations of RDF required for efficient reset operations is sufficient to corrupt efficient set operations since the presence of RDF is inhibitory in the integration reaction. It was concluded that the model is capable of identifying the appropriate ranges of recombinase expression and degradation rates required of a reliable set-hold-reset-hold operative cycle [Bonnet et al., 2012]. That said, assembling the appropriate genetic constructs to realise this predicted functionality proved to be particularly challenging due to difficulties associated with inducing the expression of recombinase proteins and the timing of such induction events within *E. coli*. Establishing a functional RAD module was eventually achieved through an *ad hoc* approach involving  $\sim 400$  trials, reiterating the aforementioned experimental limitations as well as the ramifications of noise and stochastic biological processes. Given the uncertainty surrounding particular aspects of the system, such as reversibility of the excision reaction and the search for the conditions necessary for optimal RAD module operation, it is clear that modelling DNA recombination warrants further consideration.

### 3.2 Formulating a mechanistic model of *in vitro* RAD module dynamics

An extensive review of the experimental literature was carried out in order to synthesise the current state of knowledge regarding the mechanistic basis of DNA recombination. The literature review identified several mechanistic properties of the system that are well established; the application of mass action kinetics to each of the associated biochemical equations allows us to derive the system of ODEs comprising our mechanistic model, as described below.

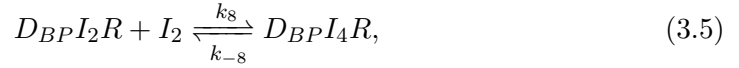
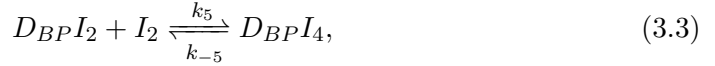
Monomeric integrase is known to form dimers reversibly in solution [Bonnet et al., 2012; Brown et al., 2011; Fogg et al., 2014; Ghosh et al., 2005, 2008; Groth and Calos, 2004; Gupta et al., 2007; Keenholtz et al., 2011; Khaleel et al., 2011; Liu et al., 2010; Lucet et al., 2005; Mandali et al., 2013; McEwan et al., 2009, 2011; Miura et al., 2011; Olorunniji and Stark, 2010; Olorunniji et al., 2012; Rowley and Smith, 2008; Singh et al., 2013; Smith et al., 2004, 2010; Stark et al., 2011; Thorpe et al., 2000; Yuan et al., 2008; Zhang et al., 2010], and thus we can write the biochemical equation:



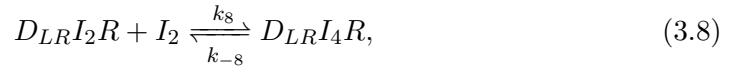
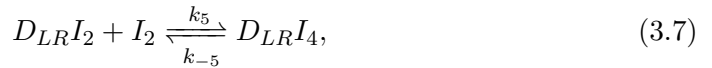
where  $I$  and  $I_2$  denote monomeric and dimeric integrase respectively and  $k_i$  are the relevant reaction rate constants. Reversible reactions are denoted using right and left arrows with the corresponding forward and backward reaction rates written above and below respectively.

One integrase dimer bound to attB and attP is necessary to mediate the integration reaction [Bonnet et al., 2012; Bai et al., 2011; Bibb et al., 2005; Breuner et al., 1999, 2001; Brown et al., 2011; Cho et al., 2002; Combes et al., 2002; Fogg et al., 2014; Ghosh et al., 2005, 2008; Groth and Calos, 2004; Gupta et al., 2007; Keenholtz et al., 2011; Keravala et al., 2006; Khaleel et al., 2011; Kim et al., 2003; Lewis and Hatfull, 2000; Liu et al., 2010; Mandali et al., 2013; Matsuura et al., 1996; McEwan et al., 2009, 2011; Miura et al., 2011; Nkrumah et al., 2006; Olivares et al., 2001; Olorunniji et al., 2012; Pena et al., 1999; Rowley and Smith, 2008; Rutherford et al., 2013; Singh et al., 2013; Smith et al., 2004, 2010; Stark et al., 2011; Thomson and Ow, 2006; Thorpe et al., 2000; Thyagarajan et al., 2001; van Duyn and Rutherford, 2013; Yuan et al., 2008; Zhang et al., 2008, 2010], which is unidirectional (irreversible) [Bonnet et al., 2012; Bai et al., 2011; Bibb and Hatfull, 2002; Bibb et al., 2005; Brown et al., 2011; Combes et al., 2002; Fogg et al., 2014;

Groth and Calos, 2004; Gupta et al., 2007; Keravala et al., 2006; Khaleel et al., 2011; Lewis and Hatfull, 2000; McEwan et al., 2009; Miura et al., 2011; Olivares et al., 2001; Olorunniji et al., 2012; Rashel et al., 2008; Rowley and Smith, 2008; Rutherford et al., 2013; Singh et al., 2013; Smith et al., 2010; Swalla et al., 2003; Thomson and Ow, 2006; Thyagarajan et al., 2001; van Duyne and Rutherford, 2013; Yuan et al., 2008; Zhang et al., 2008]. This gives the following additional dynamics:

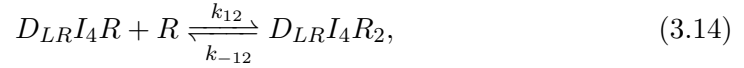
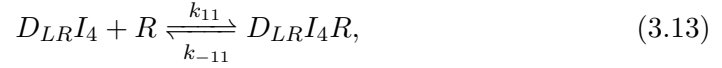
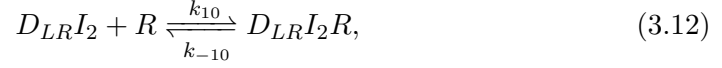
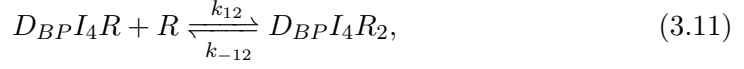
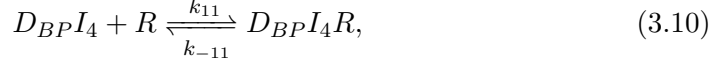
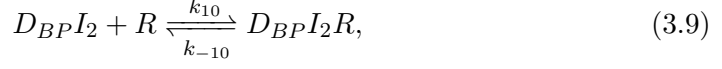


where  $D_{BP}$  denotes free DNA in the BP state;  $D_{BP}I_2/D_{BP}I_4$  denote DNA:protein complexes with one/two integrase dimers bound in the BP state respectively;  $D_{LR}I_4$  denotes the DNA:protein complex with two integrase dimers bound in the LR state;  $R$  denotes monomeric RDF (gp3) and  $D_{BP}I_2R/D_{BP}I_4R$  denote DNA:protein complexes with one/two integrase dimers and one gp3 monomer bound respectively in the BP state. Reversible reactions are denoted by double arrows with the corresponding reaction rate constants written above and below. Irreversible reactions are denoted by a single right arrow with the corresponding reaction rate constant written above. We include (3.5) here to acknowledge that dimeric integrase is able to bind to any unoccupied attachment site, however, these particular complexes are not directly involved in the integration reaction. The same dimeric integrase binding occurs in the LR state and hence gives rise to the following equivalent biochemical equations:



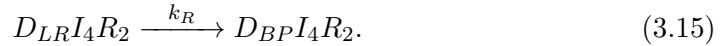
Gp3 binds to dimeric integrase already bound to DNA attachment sites and it does not bind directly to DNA attachment sites [Bonnet et al., 2012; Ghosh et al., 2006, 2008; Keenholtz et al., 2011; Khaleel et al., 2011; Singh et al., 2013; Yuan

et al., 2008]. This gives:



where  $D_{LR}I_2/D_{LR}I_4$  denote DNA:protein complexes with one/two integrase dimers bound in the LR state;  $D_{LR}I_2R/D_{LR}I_4R$  denote DNA:protein complexes with one/two integrase dimers and one gp3 monomer bound in the LR state;  $D_{BP}I_4R_2/D_{LR}I_4R_2$  denote DNA:protein complexes with two integrase dimers and two gp3 monomers bound in the BP/LR state.

Binding of gp3 to dimeric integrase already bound to both attL and attR is necessary to mediate excision, restoring attB and attP [Bonnet et al., 2012; Breuner et al., 2001; Brown et al., 2011; Combes et al., 2002; Groth and Calos, 2004; Khaleel et al., 2011; Kim et al., 2003; Lewis and Hatfull, 2000; Liu et al., 2010; Lucet et al., 2005; Mandali et al., 2013; McEwan et al., 2011; Miura et al., 2011; Nkrumah et al., 2006; Olorunniji et al., 2012; Pena et al., 1999; Singh et al., 2013; Zhang et al., 2008, 2010]:

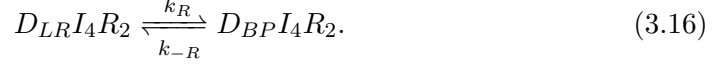


In contrast to the strong experimental evidence for each of the aforementioned mechanisms, we were unable to find a consensus in the literature regarding three further significant biological details, namely:

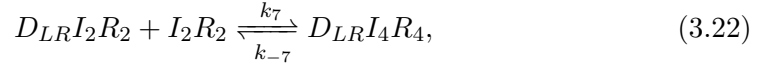
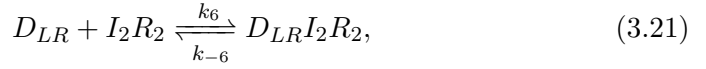
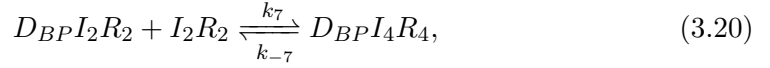
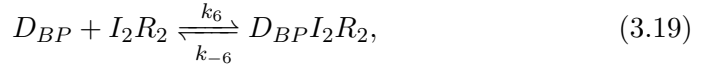
1. The directionality of the excision reaction.
2. Gp3 dimerisation and subsequent tetramerisation in solution.
3. Monomeric integrase binding to DNA substrates.

These properties all represent potentially valid mechanisms within a model of DNA recombination, each resulting in a mathematical model with distinct features. In the

case of excision reaction directionality, bidirectional excision results in the following biochemical equation:

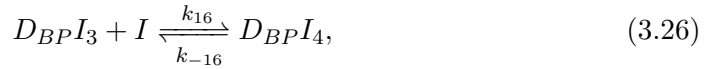
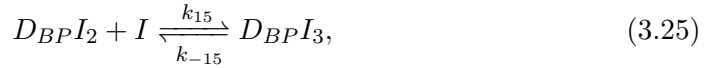
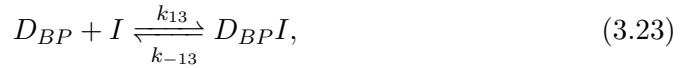


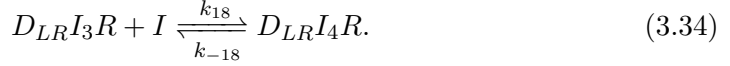
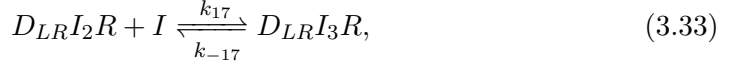
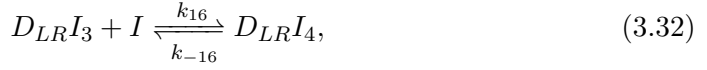
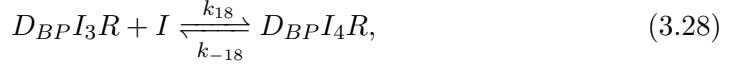
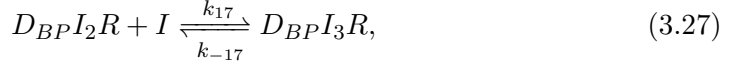
Implementing gp3 dimerisation and subsequent tetramerisation requires a significantly greater number of additional biochemical equations, since this gives rise to monomeric gp3 forming dimeric gp3, with these dimers binding to dimeric integrase to form a tetramer in solution, and this tetramer being able to bind directly to DNA attachment sites. The resultant biochemical equations are as follows:



where  $R_2$  denotes dimeric gp3;  $I_2R_2$  denotes the integrase:gp3 tetramer;  $D_{BP}I_2R_2/D_{LR}I_2R_2$  and  $D_{BP}I_4R_4/D_{LR}I_4R_4$  denote the DNA:protein complexes with one/two integrase:gp3 tetramers bound in the BP/LR state. We also investigated the performance of a model accounting for gp3 dimerisation only, with no subsequent tetramerisation in solution.

Monomeric integrase binding to DNA substrates contributes a further twelve biochemical equations due to the fact that twelve intermediate complexes arise from monomeric integrase binding compared to simplistic pairwise, dimeric binding:





We also investigated a variety of other models accounting for combinations of the aforementioned mechanisms as well as alternative gp3 binding mechanisms. Initial testing of all potential models revealed a consistently higher recombination efficiency than that observed in our experimental data. Given that *in vitro* integrase dimerisation can potentially result in the formation of dysfunctional dimers that are unable to bind effectively to DNA attachment sites, we incorporated a mechanism whereby dysfunctional integrase dimers,  $I_{2X}$ , form irreversibly in addition to the reversible formation of functional dimeric integrase:



where  $k_{\text{int}X}$  denotes the rate of dysfunctional integrase dimerisation. As expected, this mechanism reduced the concentration of functional dimeric integrase,  $I_2$ , and hence overall recombination efficiency, since integrase is involved in the mediation of both recombination reactions.

We tested different models capturing alternative mechanisms implementing each of the above features against our experimental data (see following section). The model structure which showed the capability to best match the data is depicted in Figure 3.3. Our optimal reaction network consists of a unidirectional excision reaction and monomeric integrase binding, and includes the formation of dysfunctional integrase dimers and 2:1 integrase:RDF stoichiometry of the synaptic complexes. When versions of the model incorporating gp3 dimerisation and subsequent tetramerisation in solution were optimised against the experimental data,

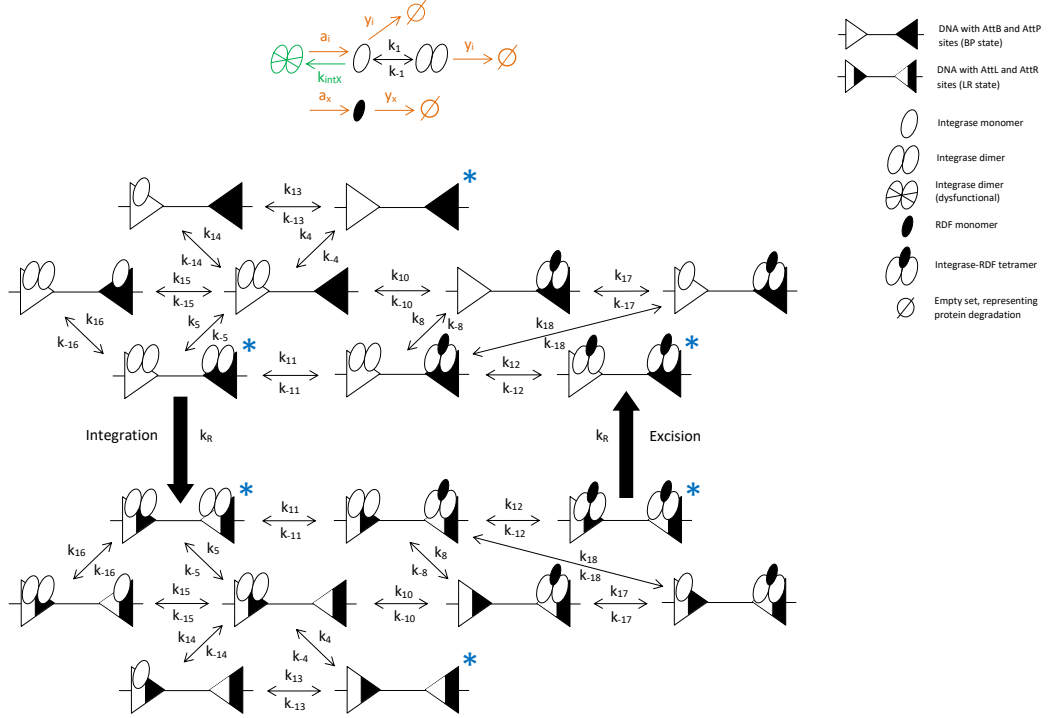


Figure 3.3: The DNA recombination reaction network used to derive both our *in vitro* and *in vivo* mechanistic models. Molecular entities, reactions and reaction rate constants common to both models are depicted in black; those depicted in green describe the *in vitro* network and those depicted in orange describe the *in vivo* network. Blue asterisks highlight DNA:protein complexes that are present in the model of Bonnet et al. [2012], with the exception of the number of RDF monomers required to mediate excision. The networks are based on the mechanisms that have been verified in the current experimental literature along with others validated computationally. We model the dynamics of  $\phi$ C31 integrase and its RDF, gp3. Reactions and their rate constants are depicted by arrows and their corresponding numbered  $k$ . The rate of recombination is denoted by  $k_R$ . Figure adapted from Bowyer et al. [2015].

the minimum error observed between the data and model outputs was larger than that observed for models that do not account for the same mechanisms. We therefore omitted these mechanisms from the model. The application of mass action kinetics to the biochemical equations identified through literature mining derives the following system of 22 model ODEs:

$$\begin{aligned}
\frac{d[I]}{dt} = & 2k_{-1}[I_2] - 2k_1[I]^2 - 2k_{\text{intX}}[I]^2 + \\
& k_{-13}[D_{BP}I] - k_{13}[D_{BP}][I] + k_{-13}[D_{LR}I] - k_{13}[D_{LR}][I] + \\
& k_{-14}[D_{BP}I_2] - k_{14}[D_{BP}I][I] + k_{-14}[D_{LR}I_2] - k_{14}[D_{LR}I][I] + \\
& k_{-15}[D_{BP}I_3] - k_{15}[D_{BP}I_2][I] + k_{-15}[D_{LR}I_3] - k_{15}[D_{LR}I_2][I] + \\
& k_{-16}[D_{BP}I_4] - k_{16}[D_{BP}I_3][I] + k_{-16}[D_{LR}I_4] - k_{16}[D_{LR}I_3][I] + \\
& k_{-17}[D_{BP}I_3R] - k_{17}[D_{BP}I_2R][I] + k_{-17}[D_{LR}I_3R] - k_{17}[D_{LR}I_2R][I] + \\
& k_{-18}[D_{BP}I_4R] - k_{18}[D_{BP}I_3R][I] + k_{-18}[D_{LR}I_4R] - k_{18}[D_{LR}I_3R][I],
\end{aligned} \tag{3.36}$$

$$\begin{aligned}
\frac{d[I_2]}{dt} = & k_1[I]^2 - k_{-1}[I_2] + \\
& k_{-4}[D_{BP}I_2] - k_4[D_{BP}][I_2] + k_{-4}[D_{LR}I_2] - k_4[D_{LR}][I_2] + \\
& k_{-5}[D_{BP}I_4] - k_5[D_{BP}I_2][I_2] + k_{-5}[D_{LR}I_4] - k_5[D_{LR}I_2][I_2] + \\
& k_{-8}[D_{BP}I_4R] - k_8[D_{BP}I_2R][I_2] + k_{-8}[D_{LR}I_4R] - k_8[D_{LR}I_2R][I_2],
\end{aligned} \tag{3.37}$$

$$\begin{aligned}
\frac{d[R]}{dt} = & k_{-10}[D_{BP}I_2R] - k_{10}[D_{BP}I_2][R] + k_{-10}[D_{LR}I_2R] - k_{10}[D_{LR}I_2][R] + \\
& k_{-11}[D_{BP}I_4R] - k_{11}[D_{BP}I_4][R] + k_{-11}[D_{LR}I_4R] - k_{11}[D_{LR}I_4][R] + \\
& k_{-12}[D_{BP}I_4R_2] - k_{12}[D_{BP}I_4R][R] + k_{-12}[D_{LR}I_4R_2] - k_{12}[D_{LR}I_4R][R],
\end{aligned} \tag{3.38}$$

$$\frac{d[D_{BP}]}{dt} = k_{-4}[D_{BP}I_2] - k_4[D_{BP}][I_2] + k_{-13}[D_{BP}I] - k_{13}[D_{BP}][I], \tag{3.39}$$

$$\begin{aligned}
\frac{d[D_{BP}I_2]}{dt} = & k_4[D_{BP}][I_2] - k_{-4}[D_{BP}I_2] + k_{-5}[D_{BP}I_4] - k_5[D_{BP}I_2][I_2] + \\
& k_{-10}[D_{BP}I_2R] - k_{10}[D_{BP}I_2][R] + k_{14}[D_{BP}I][I] - k_{-14}[D_{BP}I_2] + \\
& k_{-15}[D_{BP}I_3] - k_{15}[D_{BP}I_2][I],
\end{aligned} \tag{3.40}$$

$$\begin{aligned}
\frac{d[D_{BP}I_4]}{dt} = & k_5[D_{BP}I_2][I_2] - k_{-5}[D_{BP}I_4] + k_{-11}[D_{BP}I_4R] - k_{11}[D_{BP}I_4][R] + \\
& k_{16}[D_{BP}I_3][I] - k_{-16}[D_{BP}I_4] - k_R[D_{BP}I_4],
\end{aligned} \tag{3.41}$$

$$\begin{aligned}
\frac{d[D_{BP}I_2R]}{dt} = & k_{-8}[D_{BP}I_4R] - k_8[D_{BP}I_2R][I_2] + k_{10}[D_{BP}I_2][R] - k_{-10}[D_{BP}I_2R] + \\
& k_{-17}[D_{BP}I_3R] - k_{17}[D_{BP}I_2R][I],
\end{aligned} \tag{3.42}$$

$$\begin{aligned}
\frac{d[D_{BP}I_4R]}{dt} = & k_8[D_{BP}I_2R][I_2] - k_{-8}[D_{BP}I_4R] + k_{11}[D_{BP}I_4][R] - k_{-11}[D_{BP}I_4R] + \\
& k_{-12}[D_{BP}I_4R_2] - k_{12}[D_{BP}I_4R][R] + k_{18}[D_{BP}I_3R][I] - k_{-18}[D_{BP}I_4R],
\end{aligned} \tag{3.43}$$

$$\frac{d[D_{BP}I_4R_2]}{dt} = k_{12}[D_{BP}I_4R][R] - k_{-12}[D_{BP}I_4R_2] + k_R[D_{LR}I_4R_2], \tag{3.44}$$

$$\frac{d[D_{LR}]}{dt} = k_{-4}[D_{LR}I_2] - k_4[D_{LR}][I_2] + k_{-13}[D_{LR}I] - k_{13}[D_{LR}][I], \tag{3.45}$$

$$\begin{aligned}
\frac{d[D_{LR}I_2]}{dt} = & k_4[D_{LR}][I_2] - k_{-4}[D_{LR}I_2] + k_{-5}[D_{LR}I_4] - k_5[D_{LR}I_2][I_2] + \\
& k_{-10}[D_{LR}I_2R] - k_{10}[D_{LR}I_2][R] + k_{14}[D_{LR}I][I] - k_{-14}[D_{LR}I_2] + \\
& k_{-15}[D_{LR}I_3] - k_{15}[D_{LR}I_2][I],
\end{aligned} \tag{3.46}$$

$$\begin{aligned}
\frac{d[D_{LR}I_4]}{dt} = & k_5[D_{LR}I_2][I_2] - k_{-5}[D_{LR}I_4] + k_{-11}[D_{LR}I_4R] - k_{11}[D_{LR}I_4][R] + \\
& k_{16}[D_{LR}I_3][I] - k_{-16}[D_{LR}I_4] + k_R[D_{BP}I_4],
\end{aligned} \tag{3.47}$$

$$\begin{aligned}
\frac{d[D_{LR}I_2R]}{dt} = & k_{-8}[D_{LR}I_4R] - k_8[D_{LR}I_2R][I_2] + k_{10}[D_{LR}I_2][R] - k_{-10}[D_{LR}I_2R] + \\
& k_{-17}[D_{LR}I_3R] - k_{17}[D_{LR}I_2R][I],
\end{aligned} \tag{3.48}$$



$$\begin{aligned} \frac{d[D_{LR}I_4R]}{dt} = & k_8[D_{LR}I_2R][I_2] - k_{-8}[D_{LR}I_4R] + k_{11}[D_{LR}I_4][R] - k_{-11}[D_{LR}I_4R] + \\ & k_{-12}[D_{LR}I_4R_2] - k_{12}[D_{LR}I_4R][R] + k_{18}[D_{LR}I_3R][I] - k_{-18}[D_{LR}I_4R], \end{aligned} \quad (3.49)$$

$$\frac{d[D_{LR}I_4R_2]}{dt} = k_{12}[D_{LR}I_4R][R] - k_{-12}[D_{LR}I_4R_2] - k_R[D_{LR}I_4R_2], \quad (3.50)$$

$$\frac{d[I_{2X}]}{dt} = k_{\text{intX}}[I]^2, \quad (3.51)$$

$$\frac{d[D_{BP}I]}{dt} = k_{13}[D_{BP}][I] - k_{-13}[D_{BP}I] - k_{14}[D_{BP}I][I] + k_{-14}[D_{BP}I_2], \quad (3.52)$$

$$\frac{d[D_{BP}I_3]}{dt} = k_{15}[D_{BP}I_2][I] - k_{-15}[D_{BP}I_3] - k_{16}[D_{BP}I_3][I] + k_{-16}[D_{BP}I_4], \quad (3.53)$$

$$\frac{d[D_{BP}I_3R]}{dt} = k_{17}[D_{BP}I_2R][I] - k_{-17}[D_{BP}I_3R] - k_{18}[D_{BP}I_3R][I] + k_{-18}[D_{BP}I_4R], \quad (3.54)$$

$$\frac{d[D_{LR}I]}{dt} = k_{13}[D_{LR}][I] - k_{-13}[D_{LR}I] - k_{14}[D_{LR}I][I] + k_{-14}[D_{LR}I_2], \quad (3.55)$$

$$\frac{d[D_{LR}I_3]}{dt} = k_{15}[D_{LR}I_2][I] - k_{-15}[D_{LR}I_3] - k_{16}[D_{LR}I_3][I] + k_{-16}[D_{LR}I_4], \quad (3.56)$$

$$\frac{d[D_{LR}I_3R]}{dt} = k_{17}[D_{LR}I_2R][I] - k_{-17}[D_{LR}I_3R] - k_{18}[D_{LR}I_3R][I] + k_{-18}[D_{LR}I_4R], \quad (3.57)$$

where the square bracket notation denotes concentration and the reaction rate constants form the 28 corresponding model parameters, denoted by each numbered  $k$ . The efficiency of the RAD module to switch from one DNA state to the other is taken to be the concentration of free DNA and DNA complexes in the final state as a percentage of the concentration of free DNA in the initial state. That is, we analyse the total register of the system in the DNA state of interest. This simply involves summing the ODEs corresponding to all DNA-based molecular entities of the same DNA state. Summing each set of nine ODEs corresponding to each DNA state ((3.39)-(3.44) and (3.52)-(3.54) for the BP state; (3.45)-(3.50) and (3.55)-(3.57) for the LR state) gives two ODEs describing the dynamics of the total register of the system in BP state,  $D_{BPtot}$ , and LR state,  $D_{LRtot}$ :

$$\frac{dD_{BPtot}}{dt} = -k_R D_{BP}I_4 + k_R D_{LR}I_4R_2, \quad (3.58)$$

$$\frac{dD_{LRtot}}{dt} = k_R D_{BP}I_4 - k_R D_{LR}I_4R_2. \quad (3.59)$$

Our computational model simulations are the numerical solutions to (3.58) and (3.59) and are converted to a percentage of the initial concentration of DNA to demonstrate switching efficiency. The total DNA register is calculated in this fashion for all versions of the recombination network tested. In each case, the full system of ODEs is solved numerically in order to determine the total register, that is, no attempt is made to reduce the complexity and dimensionality of our mechanistic models.

### 3.3 Model validation via global optimisation

We compared our model to the existing model of *in vivo* DNA recombination proposed in Bonnet et al. [2012] for its ability to match and predict a new set of *in vitro* data on dynamic and steady-state recombination efficiency in the presence of different concentrations of integrase and gp3. The model in Bonnet et al. [2012] is derived from a simple reaction network comprised of nine molecular entities and is void of considerable mechanistic detail such as monomeric integrase binding, the intermediate complexes arising from individual dimeric integrase and monomeric gp3 binding, and a 2:1 integrase:gp3 complex stoichiometry. To ensure the validity of the model comparisons, we adapted the model of Bonnet et al. [2012] to the *in vitro* context and also imposed the same dysfunctional integrase dimerisation mechanism from our optimal model (Figure 3.4). Optimal model performance was evaluated by using the

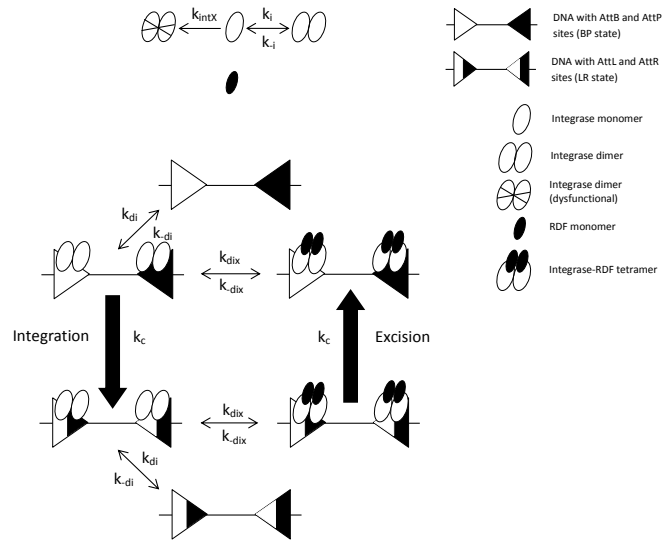


Figure 3.4: The DNA recombination reaction network adapted from Bonnet et al. [2012]. The exclusion of SSR expression and degradation coupled with the inclusion of dysfunctional integrase dimerisation accounts for our *in vitro* experimental conditions.

GA to minimise an error function defined to capture the difference between model outputs and our experimental data on *in vitro* steady-state recombination efficiency for both integration and excision reactions.

Six distinct initial concentrations (0, 50, 100, 200, 400 and 800 nM) of integrase and gp3 give the 36 pairs of initial concentrations used experimentally to record the steady-state recombination efficiency of the system. This was performed

for both the integration and excision reactions, giving an experimental dataset of 72 values. The efficiencies are given as a percentage of the initial concentration of free DNA which was set at 10 nM. Time course data was also available whereby the recombination efficiency of both reactions was recorded at 10 time points (0, 1, 2, 4, 8, 16, 32, 64, 128 and 180 minutes) for two distinct pairs of initial concentrations of integrase and gp3 (800 nM integrase, 0 nM gp3 and 400 nM integrase, 0 nM gp3). Our model is well suited to simulating this type of data since the dimensional outputs of the system (nanomolar concentrations) align with our mass action mathematical derivation, and we are able to simulate both the steady-state endpoint and full dynamical time course response of the total register over the same duration that the data were recorded (3 hours). Since our dataset was relatively large, we supplied our GA error function with a subset of the data, with the remaining data used to evaluate the predictive capabilities of our models. The subset used in data fitting was chosen to capture the full spectrum of recombination efficiencies, and all models were optimised against the same subset of experimental data and within the same parameter space.

The model of Bonnet et al. [2012] is unable to accurately match the subset of steady-state data (Figure 3.5A). In the case of the integration (BP-LR) reaction, simulations appear to be accurate for the relatively low concentration of integrase (50 nM) however, as the concentration of integrase increases, accurate fits can only be found for 800 nM integrase, 0 nM gp3. In fact, the simple model is only capable of simulating negligible recombination efficiencies for non-zero concentrations of gp3, regarding the integration reaction. This may appear to be an intuitive result given that integrase alone mediates integration, however, our data clearly indicates that the system can achieve high integration efficiencies in the presence of both SSRs. Similarly for the excision (LR-BP) reaction, simulations appear to be accurate for 50 nM integrase, but are unable to match the majority of data as integrase concentration increases. In fact, the model is only capable of simulating negligible recombination efficiencies for 0 nM gp3 and uniform efficiencies for all non-zero gp3 concentrations, regarding the excision reaction. The former observation is intuitive, since gp3 is required in combination with integrase to mediate excision, but we have no logical reason to justify the latter. With regards to prediction, we observe the same trends for both reactions meaning that the model of Bonnet et al. [2012] is void of the predictive qualities required of a useful design tool.

In contrast, our mechanistic model clearly provides a strong fit to the subset of steady-state data used in the GA global optimisation (Figure 3.5B). In the case of both reactions, we observe accurate replication of the majority of data values.

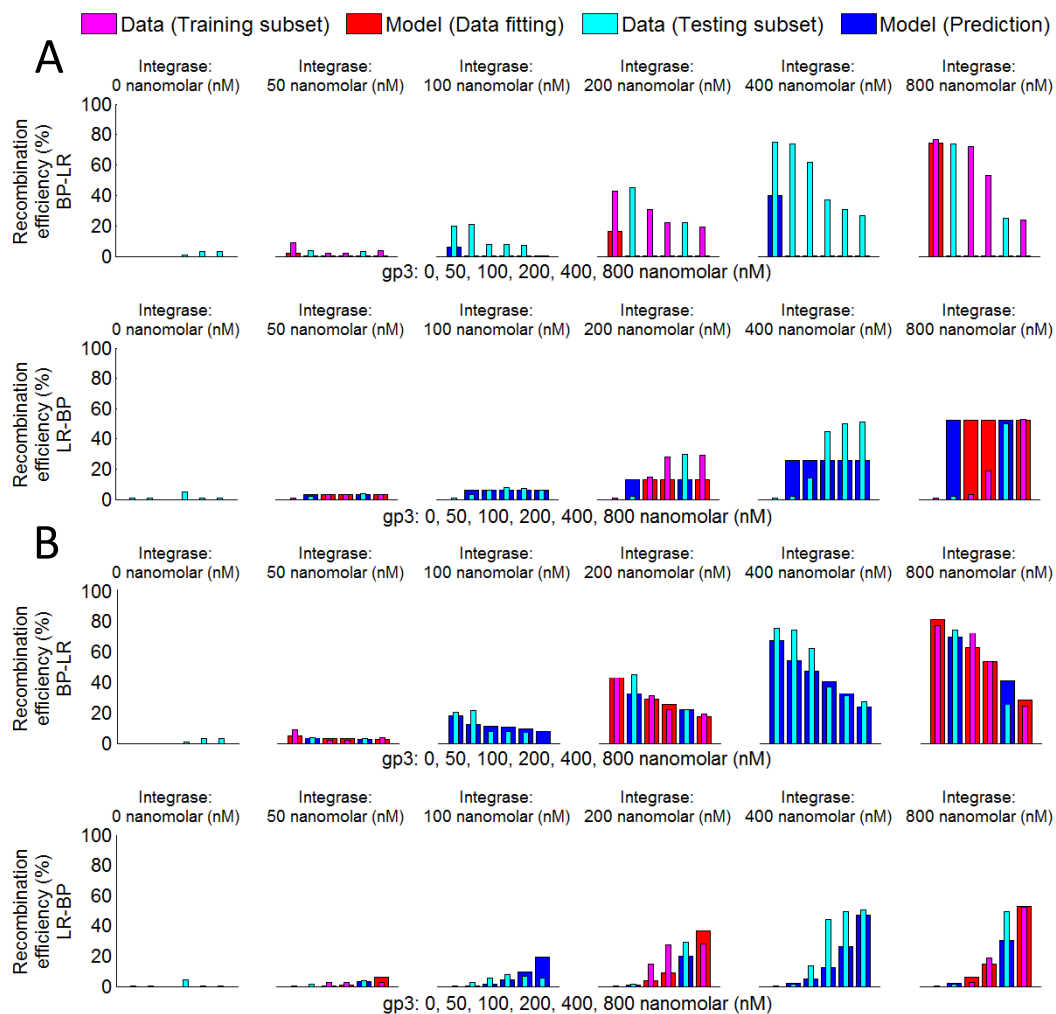


Figure 3.5: A) Data fitting/prediction results for the model of Bonnet et al. [2012]. B) Data fitting/prediction results for our mechanistic model. In both A and B the top row of bar graphs represents the integration reaction and the bottom row of bar graphs represents the excision reaction. The wider bars represent model simulations and the thinner bars represent data. Figure adapted from Bowyer et al. [2015].

The model also predicts the remaining data values effectively, presenting a clear validation of the mechanistic structure we have developed and the potential power of our model as a design tool. Similar trends are observed when we compare time course integration simulations (Figure 3.6). For 800 nM integrase, 0 nM gp3 the model of Bonnet et al. [2012] replicates the overall efficiency, but does not perform to the same extent for 400 nM integrase, 0 nM gp3. In both cases, the model exhibits a step-like response which does not match the gradual response recorded

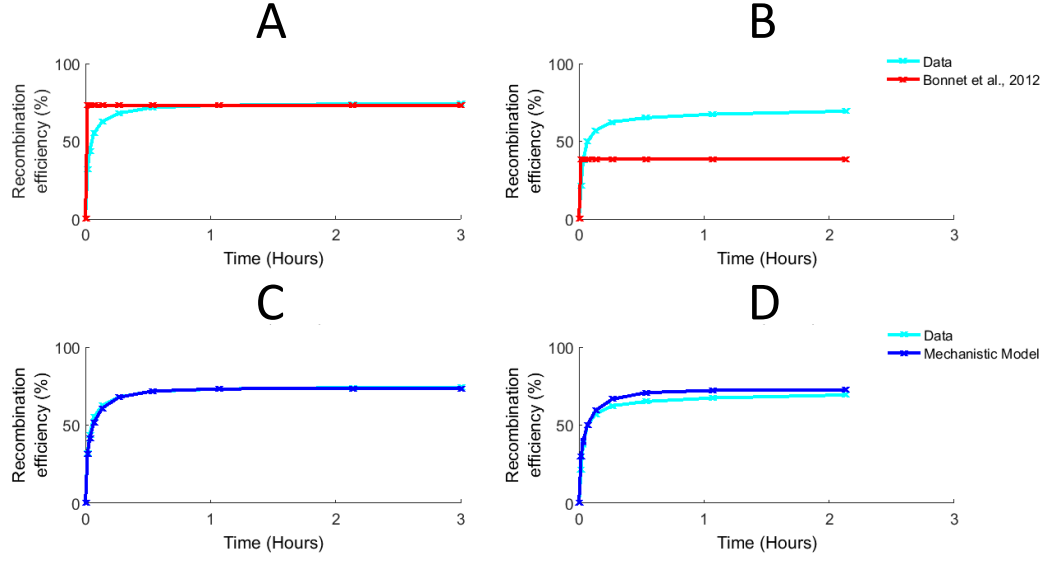


Figure 3.6: Data fitting results for the model of Bonnet et al. [2012] and our mechanistic model against time course data. Model simulations are plotted against two integration reaction time course data sets; A) and C) initiated with 800 nM integrase; B) and D) initiated with 400 nM integrase. Figure adapted from Bowyer et al. [2015].

experimentally. Again, our mechanistic model shows much improved performance, capturing the final efficiencies as well as the appropriate dynamical response in both cases.

We employed the GA function in MATLAB in order to optimise our models against the experimental data. For each model the reaction rate constants  $k_i$  were chosen as optimisation variables. The GA mimics natural selection; converging to the global minimum within the allocated parameter space by evolving an initial population of randomly generated solutions over a large number of generations. The probability of obtaining the global optimum solution is maximised by selecting the largest population size and number of generations possible. However, increasing the computational workload in this manner also greatly increases the time frame required for the algorithm to converge. Establishing an effective compromise is key for successful deployment. After a number of trials, the following parameter values for the GA were chosen:

- Population: 100
- Generations: 1000

- Bounds imposed on parameter values: [0.001, 1000]

A population size of 100 was selected for the vast majority of optimisations however, in cases where the number of model parameters was significantly increased, we increased this figure to ensure that the population:parameters ratio was never less than 3:1. We selected a particularly large parameter space due to our focus on establishing optimal model performance in light of the lack of documentation regarding reaction rates. The GA runs the given model under the same conditions used experimentally, with the resulting *in silico* recombination efficiencies subtracted from the *in vitro* data values to give a matrix of error values. We then take the absolute value of each matrix entry and then calculate the total sum to give an overall error,  $E$ , specifically:

$$E = \sum_{j=1}^4 \sum_{i=1}^3 |p_{ij}| + \sum_{j=1}^4 \sum_{i=1}^3 |q_{ij}|, \quad p_{ij} \in P, \quad q_{ij} \in Q \quad (3.60)$$

where  $P = S_{BP} - D_{BP}$  and  $Q = S_{LR} - D_{LR}$  i.e.  $P$  and  $Q$  are the  $3 \times 4$  matrices calculated by subtracting the matrices of data values ( $D_{BP}$ ,  $D_{LR}$ ) from the matrices of model simulations ( $S_{BP}$ ,  $S_{LR}$ ) for the BP and LR reactions respectively. The matrices are comprised of 12 elements since our data subset contains twelve values for each reaction, and hence the model performs 12 corresponding simulations. The end result is a set of model parameters that gave rise to the minimum overall error (Table 3.1).

Repeated optimisation using identical GA options revealed similar optimal parameter sets, indicating that optimal solutions are limited to this small subset of the parameter space. On inspection, the variation in the orders of magnitude across the optimal parameter sets identified for both the simple and mechanistic models is small, and hence it is likely that these optimal parameter values are biologically plausible. That said, there is no guarantee that the rates of the reactions comprising biological systems are always relatively similar in magnitude and therefore experimental measurement will always provide the clearest indication of parameter space constraints and notions of parameter plausibility.

Although the simplified model presented in Bonnet et al. [2012] is unable to provide the quantitative simulations we require, such models can be beneficial in other respects. Simplification reduces the dimensionality and non-linearity of a model, which ultimately reduces the complexity of any subsequent mathematical analysis. Overall, decisions regarding the extent of model refinement should align with specific research objectives. Here, we prioritise the construction of models that

A			
Optimal parameters: simple model			
Reaction	Value ( $\text{nM}^{-1}\text{s}^{-1}$ )	Reaction	Value ( $\text{s}^{-1}$ )
$k_i$	0.009	$k_{-i}$	0.616
$k_{di}$	2.587	$k_{-di}$	4.140
$k_{dix}$	0.198	$k_{-dix}$	0.193
$k_{\text{intX}}$	0.334	$k_c$	0.261

B			
Optimal parameters: mechanistic model			
Reaction	Value ( $\text{nM}^{-1}\text{s}^{-1}$ )	Reaction	Value ( $\text{s}^{-1}$ )
$k_1$	163.949	$k_{-1}$	352.237
$k_4$	509.124	$k_{-4}$	953.116
$k_5$	210.936	$k_{-5}$	397.805
$k_8$	274.334	$k_{-8}$	777.991
$k_{10}$	54.654	$k_{-10}$	530.778
$k_{11}$	0.251	$k_{-11}$	321.409
$k_{12}$	0.757	$k_{-12}$	716.045
$k_{13}$	109.882	$k_{-13}$	157.660
$k_{14}$	1.501	$k_{-14}$	201.079
$k_{15}$	0.470	$k_{-15}$	354.830
$k_{16}$	477.208	$k_{-16}$	985.078
$k_{17}$	137.581	$k_{-17}$	560.996
$k_{18}$	223.514	$k_{-18}$	466.036
$k_{\text{intX}}$	533.668	$k_R$	303.965

Table 3.1: A) Optimal parameter values for the simple *in vitro* model adapted from Bonnet et al. [2012], used to generate Figure 3.5A. B) Optimal parameter values for our mechanistic *in vitro* model, used to generate Figure 3.5B. These optimal parameter sets are dimensional with reaction rates in the first and second columns taking the units  $\text{nM}^{-1}\text{s}^{-1}$  and  $\text{s}^{-1}$  respectively from standard mass action kinetics.

provide the most accurate quantitative outputs possible, and hence maximal mechanistic detail is retained at the expense of mathematical analysis. Future studies will readdress this balance in order to determine whether the fundamental mathematical properties of the system can be verified experimentally.

The improvement in performance provided by our mechanistic model, compared to the simpler model, is not parameter-dependent. Increased mechanistic detail does lead to an increased number of model parameters in this case, however simply increasing the number of parameters does not ensure greater accuracy. This is clear given that a number of mechanisms that were explored through model development resulted in an increase in the minimal error determined via global optimisation, despite the fact that their inclusion also increased the number of parameters (see section 3.2). Model selection methods, such as the ABC-SMC inference employed by ABC-SysBio, are effective in establishing the relative quality of different candidate models regardless of whether each candidate has the same number of parameters. Hence, the application of model selection to a set of candidates that each account for one mechanistic property of interest presents a more systematic approach to the

global optimisation trials we performed in this Chapter. It remains to verify that both methods achieve the same result.

### 3.3.1 Experimental methods

Experimental data was generated by our experimental collaborators at the University of Glasgow according to the following protocol:

Proteins ( $\phi$ C31 integrase and gp3) were purified as described in Smith et al. [2004]; Olorunniji et al. [2012]; Khaleel et al. [2011]. Integrase and gp3 were diluted in integrase dilution buffer [25 mM Tris·HCl (pH 7.5), 1 mM DTT, 1 M NaCl, and 50% (vol/vol) glycerol] and kept at  $-20\text{ }^{\circ}\text{C}$ . Substrate plasmids containing inverted repeat recombination sites (pZJ56off with attB and attP; pZJ56on with attR and attL) used for *in vitro* assay were prepared from *E. coli* DH5, using a plasmid miniprep kit (Qiagen). DNA concentrations were determined by measuring the absorbance at 260 nm.

In a typical reaction, premixed integrase and gp3 with ten times their final concentrations were added to a solution of substrate plasmid ( $\sim 10\text{ nM}$ ) in a reaction buffer [50 mM Tris·HCl (pH 7.5), 0.1 mM EDTA, 5 mM spermidine, and 0.1 mg/ml BSA]. Reactions were carried out at  $30\text{ }^{\circ}\text{C}$ , terminated at various time points, by heating the samples to  $80\text{ }^{\circ}\text{C}$  for 10 minutes. Samples were digested with restriction enzymes, then, treated with  $5\text{ }\mu\text{l}$  of loading buffer [25 mM Tris·HCl (pH 8.2), 20% (wt/vol) Ficoll, 0.5% sodium dodecyl sulphate, 5 mg/ml protease K, 0.25 mg/ml bromophenol blue] at  $37\text{ }^{\circ}\text{C}$  for 30 minutes prior to loading onto 1.2% (wt/vol) agarose gels. Gels were stained with ethidium bromide, destained in electroporation buffer, and photographed using Bio-Rad UV Transilluminator. Recombinant and non-recombinant DNA bands were quantified using the volume analysis tool of Quantity One software.

## 3.4 Non-dimensional simulations of *in vivo* RAD module dynamics

Having validated our mechanistic model against *in vitro* data, we sought to analyse model performance within an *in vivo* context. All available experimental data suggests that the DNA:protein binding interactions of the *in vitro* system are retained *in vivo*, with the introduction of protein expression and degradation representing the key adaptations. Thus, the mechanisms removed from the model of Bonnet et al. [2012] in order to analyse *in vitro* dynamics were restored and we adapted our



own mechanistic model accordingly (Figure 3.3).

Increased expression and degradation rates of the SSRs are induced to realise desired RAD module operations and hence we also account for basal expression and degradation rates occurring in the absence of induction. SSR induction is performed experimentally through chemical stimuli. The formation of dysfunctional dimeric integrase is removed from both models as we have no reason to justify its existence in this context. This reduces the number of ODEs in our mechanistic model to twenty-one since  $I_{2X}$  is no longer present:

$$\begin{aligned} \frac{d[I]}{dt} = & a_i - y_i[I] + 2k_{-1}[I_2] - 2k_1[I]^2 + \\ & k_{-13}[D_{BP}I] - k_{13}[D_{BP}][I] + k_{-13}[D_{LR}I] - k_{13}[D_{LR}][I] + \\ & k_{-14}[D_{BP}I_2] - k_{14}[D_{BP}I][I] + k_{-14}[D_{LR}I_2] - k_{14}[D_{LR}I][I] + \\ & k_{-15}[D_{BP}I_3] - k_{15}[D_{BP}I_2][I] + k_{-15}[D_{LR}I_3] - k_{15}[D_{LR}I_2][I] + \\ & k_{-16}[D_{BP}I_4] - k_{16}[D_{BP}I_3][I] + k_{-16}[D_{LR}I_4] - k_{16}[D_{LR}I_3][I] + \\ & k_{-17}[D_{BP}I_3R] - k_{17}[D_{BP}I_2R][I] + k_{-17}[D_{LR}I_3R] - k_{17}[D_{LR}I_2R][I] + \\ & k_{-18}[D_{BP}I_4R] - k_{18}[D_{BP}I_3R][I] + k_{-18}[D_{LR}I_4R] - k_{18}[D_{LR}I_3R][I], \end{aligned} \quad (3.61)$$

$$\begin{aligned} \frac{d[I_2]}{dt} = & k_1[I]^2 - k_{-1}[I_2] - y_i[I_2] + \\ & k_{-4}[D_{BP}I_2] - k_4[D_{BP}][I_2] + k_{-4}[D_{LR}I_2] - k_4[D_{LR}][I_2] + \\ & k_{-5}[D_{BP}I_4] - k_5[D_{BP}I_2][I_2] + k_{-5}[D_{LR}I_4] - k_5[D_{LR}I_2][I_2] + \\ & k_{-8}[D_{BP}I_4R] - k_8[D_{BP}I_2R][I_2] + k_{-8}[D_{LR}I_4R] - k_8[D_{LR}I_2R][I_2], \end{aligned} \quad (3.62)$$

$$\begin{aligned} \frac{d[R]}{dt} = & a_x - y_x[R] + k_{-10}[D_{BP}I_2R] - k_{10}[D_{BP}I_2][R] + k_{-10}[D_{LR}I_2R] - k_{10}[D_{LR}I_2][R] + \\ & k_{-11}[D_{BP}I_4R] - k_{11}[D_{BP}I_4][R] + k_{-11}[D_{LR}I_4R] - k_{11}[D_{LR}I_4][R] + \\ & k_{-12}[D_{BP}I_4R_2] - k_{12}[D_{BP}I_4R][R] + k_{-12}[D_{LR}I_4R_2] - k_{12}[D_{LR}I_4R][R], \end{aligned} \quad (3.63)$$

$$\frac{d[D_{BP}]}{dt} = k_{-4}[D_{BP}I_2] - k_4[D_{BP}][I_2] + k_{-13}[D_{BP}I] - k_{13}[D_{BP}][I], \quad (3.64)$$

$$\begin{aligned} \frac{d[D_{BP}I_2]}{dt} = & k_4[D_{BP}][I_2] - k_{-4}[D_{BP}I_2] + k_{-5}[D_{BP}I_4] - k_5[D_{BP}I_2][I_2] + \\ & k_{-10}[D_{BP}I_2R] - k_{10}[D_{BP}I_2][R] + k_{14}[D_{BP}I][I] - k_{-14}[D_{BP}I_2] + \\ & k_{-15}[D_{BP}I_3] - k_{15}[D_{BP}I_2][I], \end{aligned} \quad (3.65)$$

$$\begin{aligned} \frac{d[D_{BP}I_4]}{dt} = & k_5[D_{BP}I_2][I_2] - k_{-5}[D_{BP}I_4] + k_{-11}[D_{BP}I_4R] - k_{11}[D_{BP}I_4][R] + \\ & k_{16}[D_{BP}I_3][I] - k_{-16}[D_{BP}I_4] - k_R[D_{BP}I_4], \end{aligned} \quad (3.66)$$

$$\begin{aligned} \frac{d[D_{BP}I_2R]}{dt} = & k_{-8}[D_{BP}I_4R] - k_8[D_{BP}I_2R][I_2] + k_{10}[D_{BP}I_2][R] - k_{-10}[D_{BP}I_2R] + \\ & k_{-17}[D_{BP}I_3R] - k_{17}[D_{BP}I_2R][I], \end{aligned} \quad (3.67)$$

$$\begin{aligned} \frac{d[D_{BP}I_4R]}{dt} = & k_8[D_{BP}I_2R][I_2] - k_{-8}[D_{BP}I_4R] + k_{11}[D_{BP}I_4][R] - k_{-11}[D_{BP}I_4R] + \\ & k_{-12}[D_{BP}I_4R_2] - k_{12}[D_{BP}I_4R][R] + k_{18}[D_{BP}I_3R][I] - k_{-18}[D_{BP}I_4R], \end{aligned} \quad (3.68)$$

$$\frac{d[D_{BP}I_4R_2]}{dt} = k_{12}[D_{BP}I_4R][R] - k_{-12}[D_{BP}I_4R_2] + k_R[D_{LR}I_4R_2], \quad (3.69)$$

$$\frac{d[D_{LR}]}{dt} = k_{-4}[D_{LR}I_2] - k_4[D_{LR}][I_2] + k_{-13}[D_{LR}I] - k_{13}[D_{LR}][I], \quad (3.70)$$

$$\begin{aligned} \frac{d[D_{LR}I_2]}{dt} = & k_4[D_{LR}][I_2] - k_{-4}[D_{LR}I_2] + k_{-5}[D_{LR}I_4] - k_5[D_{LR}I_2][I_2] + \\ & k_{-10}[D_{LR}I_2R] - k_{10}[D_{LR}I_2][R] + k_{14}[D_{LR}I][I] - k_{-14}[D_{LR}I_2] + \\ & k_{-15}[D_{LR}I_3] - k_{15}[D_{LR}I_2][I], \end{aligned} \quad (3.71)$$

$$\begin{aligned} \frac{d[D_{LR}I_4]}{dt} = & k_5[D_{LR}I_2][I_2] - k_{-5}[D_{LR}I_4] + k_{-11}[D_{LR}I_4R] - k_{11}[D_{LR}I_4][R] + \\ & k_{16}[D_{LR}I_3][I] - k_{-16}[D_{LR}I_4] + k_R[D_{BP}I_4], \end{aligned} \quad (3.72)$$

$$\begin{aligned} \frac{d[D_{LR}I_2R]}{dt} = & k_{-8}[D_{LR}I_4R] - k_8[D_{LR}I_2R][I_2] + k_{10}[D_{LR}I_2][R] - k_{-10}[D_{LR}I_2R] + \\ & k_{-17}[D_{LR}I_3R] - k_{17}[D_{LR}I_2R][I], \end{aligned} \quad (3.73)$$

$$\begin{aligned} \frac{d[D_{LR}I_4R]}{dt} = & k_8[D_{LR}I_2R][I_2] - k_{-8}[D_{LR}I_4R] + k_{11}[D_{LR}I_4][R] - k_{-11}[D_{LR}I_4R] + \\ & k_{-12}[D_{LR}I_4R_2] - k_{12}[D_{LR}I_4R][R] + k_{18}[D_{LR}I_3R][I] - k_{-18}[D_{LR}I_4R], \end{aligned} \quad (3.74)$$

$$\frac{d[D_{LR}I_4R_2]}{dt} = k_{12}[D_{LR}I_4R][R] - k_{-12}[D_{LR}I_4R_2] - k_R[D_{LR}I_4R_2], \quad (3.75)$$

$$\frac{d[D_{BP}I]}{dt} = k_{13}[D_{BP}][I] - k_{-13}[D_{BP}I] - k_{14}[D_{BP}I][I] + k_{-14}[D_{BP}I_2], \quad (3.76)$$

$$\frac{d[D_{BP}I_3]}{dt} = k_{15}[D_{BP}I_2][I] - k_{-15}[D_{BP}I_3] - k_{16}[D_{BP}I_3][I] + k_{-16}[D_{BP}I_4], \quad (3.77)$$

$$\frac{d[D_{BP}I_3R]}{dt} = k_{17}[D_{BP}I_2R][I] - k_{-17}[D_{BP}I_3R] - k_{18}[D_{BP}I_3R][I] + k_{-18}[D_{BP}I_4R], \quad (3.78)$$

$$\frac{d[D_{LR}I]}{dt} = k_{13}[D_{LR}][I] - k_{-13}[D_{LR}I] - k_{14}[D_{LR}I][I] + k_{-14}[D_{LR}I_2], \quad (3.79)$$

$$\frac{d[D_{LR}I_3]}{dt} = k_{15}[D_{LR}I_2][I] - k_{-15}[D_{LR}I_3] - k_{16}[D_{LR}I_3][I] + k_{-16}[D_{LR}I_4], \quad (3.80)$$

$$\frac{d[D_{LR}I_3R]}{dt} = k_{17}[D_{LR}I_2R][I] - k_{-17}[D_{LR}I_3R] - k_{18}[D_{LR}I_3R][I] + k_{-18}[D_{LR}I_4R], \quad (3.81)$$

where  $a_i$  and  $y_i$  denote the expression and degradation rates of integrase respectively and  $a_x$  and  $y_x$  denote the expression and degradation rates of RDF respectively. All *in vivo* modelling utilises the model of Bonnet et al. [2012] and our mechanistic model developed previously in their non-dimensional forms, in order to permit valid numerical simulations and direct mathematical comparisons [Bonnet et al., 2012; Wu et al., 2013]. Analysis of the dimensionality of the terms in our mechanistic model reveals that our model parameters have one of three dimensions:

$$[a_i] = [a_x] = \frac{M}{T}, \quad (3.82)$$

$$[y_i] = [y_x] = [k_{-n}] = [k_R] = \frac{1}{T}, \quad (3.83)$$

$$[k_n] = \frac{1}{MT}, \quad (3.84)$$

for  $n \in \{1, 4, 5, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18\}$ . Hence, we have,

$$\left[ \frac{k_{-1}}{k_1} \right] = \frac{[k_{-1}]}{[k_1]} = \frac{\frac{1}{T}}{\frac{1}{MT}} = \frac{1}{T} \cdot \frac{MT}{1} = M, \quad (3.85)$$

which identifies the ratio  $\frac{k_{-1}}{k_1}$  as having the appropriate dimensionality for rescaling concentration. With this, and taking the reciprocal of  $k_R$  with the appropriate dimensionality to rescale time ( $T$ ), we can introduce the following non-dimensional variables:

$$X = \frac{k_{-1}}{k_1} \hat{X} \quad \forall X \in (3.61) - (3.81), \quad (3.86)$$

$$t = \frac{1}{k_R} \tau, \quad (3.87)$$

where  $X$  represents all dependent variables in our model (equations (3.61)-(3.81)),  $\hat{X}$  represents the corresponding non-dimensional variable and  $\tau$  represents non-dimensional time. Substituting (3.86) and (3.87) into (3.61)-(3.81) yields our full non-dimensional model:

$$\begin{aligned} \frac{d\hat{I}}{d\tau} = & \bar{a}_i - \bar{y}_i \hat{I} + 2\bar{k}_{-1} \hat{I}_2 - 2\bar{k}_{-1} \hat{I}^2 + \\ & \bar{k}_{-13} D_{BP} \hat{I} - \bar{k}_{13} D_{BP} \hat{I} + \bar{k}_{-13} D_{LR} \hat{I} - \bar{k}_{13} D_{LR} \hat{I} + \\ & \bar{k}_{-14} D_{BP} \hat{I}_2 - \bar{k}_{14} D_{BP} \hat{I} \hat{I} + \bar{k}_{-14} D_{LR} \hat{I}_2 - \bar{k}_{14} D_{LR} \hat{I} \hat{I} + \\ & \bar{k}_{-15} D_{BP} \hat{I}_3 - \bar{k}_{15} D_{BP} \hat{I}_2 \hat{I} + \bar{k}_{-15} D_{LR} \hat{I}_3 - \bar{k}_{15} D_{LR} \hat{I}_2 \hat{I} + \\ & \bar{k}_{-16} D_{BP} \hat{I}_4 - \bar{k}_{16} D_{BP} \hat{I}_3 \hat{I} + \bar{k}_{-16} D_{LR} \hat{I}_4 - \bar{k}_{16} D_{LR} \hat{I}_3 \hat{I} + \\ & \bar{k}_{-17} D_{BP} \hat{I}_3 R - \bar{k}_{17} D_{BP} \hat{I}_2 R \hat{I} + \bar{k}_{-17} D_{LR} \hat{I}_3 R - \bar{k}_{17} D_{LR} \hat{I}_2 R \hat{I} + \\ & \bar{k}_{-18} D_{BP} \hat{I}_4 R - \bar{k}_{18} D_{BP} \hat{I}_3 R \hat{I} + \bar{k}_{-18} D_{LR} \hat{I}_4 R - \bar{k}_{18} D_{LR} \hat{I}_3 R \hat{I}, \end{aligned} \quad (3.88)$$

$$\begin{aligned} \frac{d\hat{I}_2}{d\tau} = & \bar{k}_{-1} \hat{I}^2 - \bar{k}_{-1} \hat{I}_2 - \bar{y}_i \hat{I}_2 + \\ & \bar{k}_{-4} D_{BP} \hat{I}_2 - \bar{k}_4 D_{BP} \hat{I}_2 + \bar{k}_{-4} D_{LR} \hat{I}_2 - \bar{k}_4 D_{LR} \hat{I}_2 + \\ & \bar{k}_{-5} D_{BP} \hat{I}_4 - \bar{k}_5 D_{BP} \hat{I}_2 \hat{I}_2 + \bar{k}_{-5} D_{LR} \hat{I}_4 - \bar{k}_5 D_{LR} \hat{I}_2 \hat{I}_2 + \\ & \bar{k}_{-8} D_{BP} \hat{I}_4 R - \bar{k}_8 D_{BP} \hat{I}_2 R \hat{I}_2 + \bar{k}_{-8} D_{LR} \hat{I}_4 R - \bar{k}_8 D_{LR} \hat{I}_2 R \hat{I}_2, \end{aligned} \quad (3.89)$$

$$\begin{aligned} \frac{d\hat{R}}{d\tau} = & \bar{a}_x - \bar{y}_x \hat{R} + \bar{k}_{-10} D_{BP} \hat{I}_2 R - \bar{k}_{10} D_{BP} \hat{I}_2 \hat{R} + \bar{k}_{-10} D_{LR} \hat{I}_2 R - \bar{k}_{10} D_{LR} \hat{I}_2 \hat{R} + \\ & \bar{k}_{-11} D_{BP} \hat{I}_4 R - \bar{k}_{11} D_{BP} \hat{I}_4 \hat{R} + \bar{k}_{-11} D_{LR} \hat{I}_4 R - \bar{k}_{11} D_{LR} \hat{I}_4 \hat{R} + \\ & \bar{k}_{-12} D_{BP} \hat{I}_4 R_2 - \bar{k}_{12} D_{BP} \hat{I}_4 R \hat{R} + \bar{k}_{-12} D_{LR} \hat{I}_4 R_2 - \bar{k}_{12} D_{LR} \hat{I}_4 R \hat{R}, \end{aligned} \quad (3.90)$$

$$\frac{dD_{BP}}{d\tau} = \bar{k}_{-4} D_{BP} \hat{I}_2 - \bar{k}_4 D_{BP} \hat{I}_2 + \bar{k}_{-13} D_{BP} \hat{I} - \bar{k}_{13} D_{BP} \hat{I}, \quad (3.91)$$

$$\begin{aligned} \frac{dD_{BP} \hat{I}_2}{d\tau} = & \bar{k}_4 D_{BP} \hat{I}_2 - \bar{k}_{-4} D_{BP} \hat{I}_2 + \bar{k}_{-5} D_{BP} \hat{I}_4 - \bar{k}_5 D_{BP} \hat{I}_2 \hat{I}_2 + \\ & \bar{k}_{-10} D_{BP} \hat{I}_2 R - \bar{k}_{10} D_{BP} \hat{I}_2 \hat{R} + \bar{k}_{14} D_{BP} \hat{I} \hat{I} - \bar{k}_{-14} D_{BP} \hat{I}_2 + \\ & \bar{k}_{-15} D_{BP} \hat{I}_3 - \bar{k}_{15} D_{BP} \hat{I}_2 \hat{I}, \end{aligned} \quad (3.92)$$

$$\begin{aligned} \frac{dD_{BP} \hat{I}_4}{d\tau} = & \bar{k}_5 D_{BP} \hat{I}_2 \hat{I}_2 - \bar{k}_{-5} D_{BP} \hat{I}_4 + \bar{k}_{-11} D_{BP} \hat{I}_4 R - \bar{k}_{11} D_{BP} \hat{I}_4 \hat{R} + \\ & \bar{k}_{16} D_{BP} \hat{I}_3 \hat{I} - \bar{k}_{-16} D_{BP} \hat{I}_4 - D_{BP} \hat{I}_4, \end{aligned} \quad (3.93)$$

$$\begin{aligned} \frac{dD_{BP} \hat{I}_2 R}{d\tau} = & \bar{k}_{-8} D_{BP} \hat{I}_4 R - \bar{k}_8 D_{BP} \hat{I}_2 R \hat{I}_2 + \bar{k}_{10} D_{BP} \hat{I}_2 \hat{R} - \bar{k}_{-10} D_{BP} \hat{I}_2 R + \\ & \bar{k}_{-17} D_{BP} \hat{I}_3 R - \bar{k}_{17} D_{BP} \hat{I}_2 R \hat{I}, \end{aligned} \quad (3.94)$$

$$\begin{aligned} \frac{dD_{BP}\hat{I}_4R}{d\tau} = & \bar{k}_8 D_{BP}\hat{I}_2R\hat{I}_2 - \bar{k}_{-8} D_{BP}\hat{I}_4R + \bar{k}_{11} D_{BP}\hat{I}_4\hat{R} - \bar{k}_{-11} D_{BP}\hat{I}_4R + \\ & \bar{k}_{-12} D_{BP}\hat{I}_4R_2 - \bar{k}_{12} D_{BP}\hat{I}_4R\hat{R} + \bar{k}_{18} D_{BP}\hat{I}_3R\hat{I} - \bar{k}_{-18} D_{BP}\hat{I}_4R, \end{aligned} \quad (3.95)$$

$$\frac{dD_{BP}\hat{I}_4R_2}{d\tau} = \bar{k}_{12} D_{BP}\hat{I}_4R\hat{R} - \bar{k}_{-12} D_{BP}\hat{I}_4R_2 + D_{LR}\hat{I}_4R_2, \quad (3.96)$$

$$\frac{dD_{LR}\hat{I}_2}{d\tau} = \bar{k}_{-4} D_{LR}\hat{I}_2 - \bar{k}_4 D_{LR}\hat{I}_2 + \bar{k}_{-13} D_{LR}\hat{I} - \bar{k}_{13} D_{LR}\hat{I}, \quad (3.97)$$

$$\begin{aligned} \frac{dD_{LR}\hat{I}_2}{d\tau} = & \bar{k}_4 D_{LR}\hat{I}_2 - \bar{k}_{-4} D_{LR}\hat{I}_2 + \bar{k}_{-5} D_{LR}\hat{I}_4 - \bar{k}_5 D_{LR}\hat{I}_2\hat{I}_2 + \\ & \bar{k}_{-10} D_{LR}\hat{I}_2R - \bar{k}_{10} D_{LR}\hat{I}_2\hat{R} + \bar{k}_{14} D_{LR}\hat{I}\hat{I} - \bar{k}_{-14} D_{LR}\hat{I}_2 + \\ & \bar{k}_{-15} D_{LR}\hat{I}_3 - \bar{k}_{15} D_{LR}\hat{I}_2\hat{I}, \end{aligned} \quad (3.98)$$

$$\begin{aligned} \frac{dD_{LR}\hat{I}_4}{d\tau} = & \bar{k}_5 D_{LR}\hat{I}_2\hat{I}_2 - \bar{k}_{-5} D_{LR}\hat{I}_4 + \bar{k}_{-11} D_{LR}\hat{I}_4R - \bar{k}_{11} D_{LR}\hat{I}_4\hat{R} + \\ & \bar{k}_{16} D_{LR}\hat{I}_3\hat{I} - \bar{k}_{-16} D_{LR}\hat{I}_4 + D_{BP}\hat{I}_4, \end{aligned} \quad (3.99)$$

$$\begin{aligned} \frac{dD_{LR}\hat{I}_2R}{d\tau} = & \bar{k}_{-8} D_{LR}\hat{I}_4R - \bar{k}_8 D_{LR}\hat{I}_2R\hat{I}_2 + \bar{k}_{10} D_{LR}\hat{I}_2\hat{R} - \bar{k}_{-10} D_{LR}\hat{I}_2R + \\ & \bar{k}_{-17} D_{LR}\hat{I}_3R - \bar{k}_{17} D_{LR}\hat{I}_2R\hat{I}, \end{aligned} \quad (3.100)$$

$$\begin{aligned} \frac{dD_{LR}\hat{I}_4R}{d\tau} = & \bar{k}_8 D_{LR}\hat{I}_2R\hat{I}_2 - \bar{k}_{-8} D_{LR}\hat{I}_4R + \bar{k}_{11} D_{LR}\hat{I}_4\hat{R} - \bar{k}_{-11} D_{LR}\hat{I}_4R + \\ & \bar{k}_{-12} D_{LR}\hat{I}_4R_2 - \bar{k}_{12} D_{LR}\hat{I}_4R\hat{R} + \bar{k}_{18} D_{LR}\hat{I}_3R\hat{I} - \bar{k}_{-18} D_{LR}\hat{I}_4R, \end{aligned} \quad (3.101)$$

$$\frac{dD_{LR}\hat{I}_4R_2}{d\tau} = \bar{k}_{12} D_{LR}\hat{I}_4R\hat{R} - \bar{k}_{-12} D_{LR}\hat{I}_4R_2 - D_{LR}\hat{I}_4R_2, \quad (3.102)$$

$$\frac{dD_{BP}\hat{I}}{d\tau} = \bar{k}_{13} D_{BP}\hat{I} - \bar{k}_{-13} D_{BP}\hat{I} - \bar{k}_{14} D_{BP}\hat{I}\hat{I} + \bar{k}_{-14} D_{BP}\hat{I}_2, \quad (3.103)$$

$$\frac{dD_{BP}\hat{I}_3}{d\tau} = \bar{k}_{15} D_{BP}\hat{I}_2\hat{I} - \bar{k}_{-15} D_{BP}\hat{I}_3 - \bar{k}_{16} D_{BP}\hat{I}_3\hat{I} + \bar{k}_{-16} D_{BP}\hat{I}_4, \quad (3.104)$$

$$\frac{dD_{BP}\hat{I}_3R}{d\tau} = \bar{k}_{17} D_{BP}\hat{I}_2R\hat{I} - \bar{k}_{-17} D_{BP}\hat{I}_3R - \bar{k}_{18} D_{BP}\hat{I}_3R\hat{I} + \bar{k}_{-18} D_{BP}\hat{I}_4R, \quad (3.105)$$

$$\frac{dD_{LR}\hat{I}}{d\tau} = \bar{k}_{13} D_{LR}\hat{I} - \bar{k}_{-13} D_{LR}\hat{I} - \bar{k}_{14} D_{LR}\hat{I}\hat{I} + \bar{k}_{-14} D_{LR}\hat{I}_2, \quad (3.106)$$

$$\frac{dD_{LR}\hat{I}_3}{d\tau} = \bar{k}_{15} D_{LR}\hat{I}_2\hat{I} - \bar{k}_{-15} D_{LR}\hat{I}_3 - \bar{k}_{16} D_{LR}\hat{I}_3\hat{I} + \bar{k}_{-16} D_{LR}\hat{I}_4, \quad (3.107)$$

$$\frac{dD_{LR}\hat{I}_3R}{d\tau} = \bar{k}_{17} D_{LR}\hat{I}_2R\hat{I} - \bar{k}_{-17} D_{LR}\hat{I}_3R - \bar{k}_{18} D_{LR}\hat{I}_3R\hat{I} + \bar{k}_{-18} D_{LR}\hat{I}_4R, \quad (3.108)$$

where the parameter  $k_1$  has been factored out of the system, thus reducing the total number of parameters by one, and we denote non-dimensional parameters using the same corresponding numerical subscripts as the dimensional model with the addition of the bar notation. The square bracket notation is no longer appropriate since we have removed dimensionality from the model.

The main focus of our *in vivo* investigation is the excision reaction, since the integration reaction is straightforward to elucidate; over-expression of integrase is guaranteed to induce highly efficient integration for relatively low gp3 expression since integration is mediated by integrase only. However, the excision reaction is nuanced by its mediation by both SSRs and directly influences the functionality of the RAD module as a result. A brute-force approach of over-expressing integrase and

gp3 simultaneously to achieve highly efficient excision is an intuitive notion before considering that desirable RAD module functionality will often require hold states. That is, once excision has been induced through simultaneous over-expression, it is likely that spontaneous re-integration will occur due to high residual integrase concentration and gp3 dissociation. Therefore it is not conceptually obvious how to induce highly efficient excision and then hold that state equally efficiently in the absence of induction. Figure 3.7 depicts the *in vivo* RAD module dynamics for 2.5 repeated OFF-HOLD-ON-HOLD operative cycles for both the simple model and our mechanistic model. We define RAD module operations as follows: ON is an

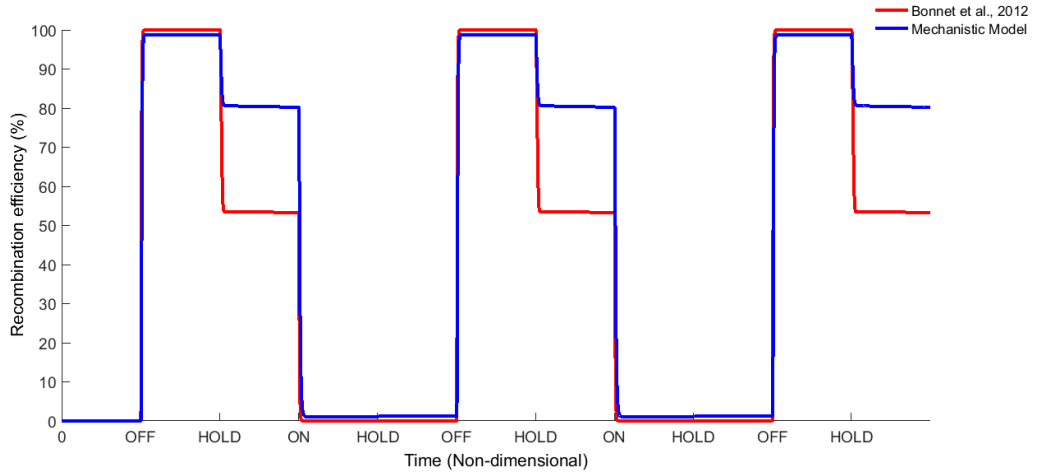


Figure 3.7: RAD module *in vivo* switching efficiencies for the model of [Bonnet et al., 2012] and our mechanistic model. Both models simulate 2.5 repeated OFF-HOLD-ON-HOLD operative cycles. All non-dimensional parameter values for both models are set equal to 1 to simulate these plots with the exception of  $a_i$  and  $a_x$ . For the ON operation  $a_i = 10$ ,  $a_x = 0.1$ ; the OFF operation  $a_i = a_x = 10$ ; the HOLD operation  $a_i = a_x = 0.1$ .

integration reaction induced through increased integrase levels only, OFF is an excision reaction induced through simultaneously increased integrase and gp3 levels and HOLD is the restoration of basal SSR levels following either ON or OFF operations. Both models exhibit consistent switching efficiencies across each of their own repeated cycles which demonstrates that the module can maintain performance over many identical induction events. That is, the process of inducing desired SSR expression/degradation should permit efficient module operations whenever required, and regardless of switching the module ON or OFF. Both models hold state efficiently following an integration reaction since basal SSR levels are restored in the absence of

induction and hence there is insufficient gp3 to mediate natural re-excision. There is, however, a notable distinction between the performance of the models in the efficiency of the HOLD state following the induction of an excision reaction. We observe natural re-integration efficiencies of  $\sim 47\%$  for the model of Bonnet et al. [2012] against  $\sim 23\%$  for our mechanistic model. This suggests that the efficiency of RAD module HOLD states following excision is, in fact, greater than expected from the model of Bonnet et al. [2012]. To determine the reasons for this, we analyse the mechanistic distinctions between the two models. Compared to the simple model, our mechanistic model accounts for the following additional mechanisms:

1. Monomeric integrase binding to free DNA substrates.
2. Formation of synaptic complexes with 2:1 stoichiometry.
3. Formation of intermediate DNA:protein complexes.

Figure 3.8 shows the effect of these biological distinctions on HOLD state efficiency following excision. We apply each distinction to the model of Bonnet et al. [2012]

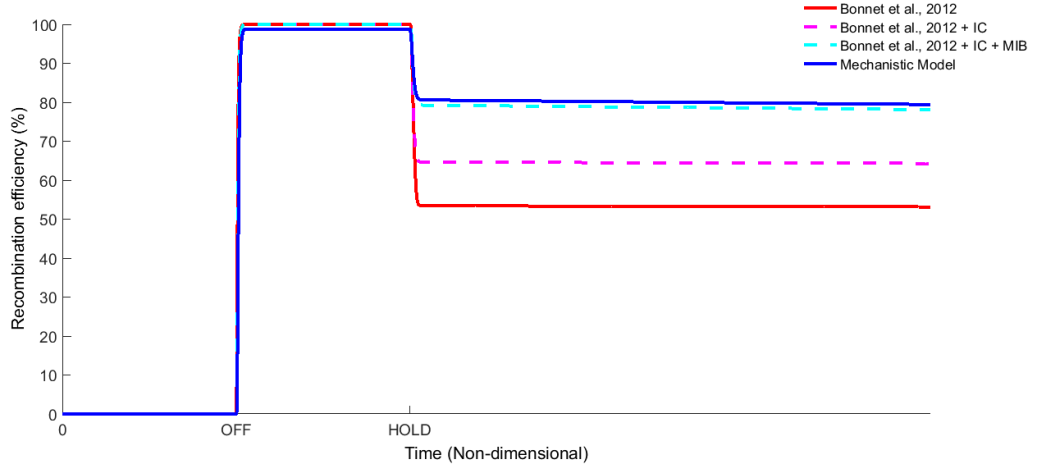


Figure 3.8: RAD module *in vivo* HOLD state efficiencies for the model of Bonnet et al. [2012] and our mechanistic model. The dashed line plots demonstrate the improvements on HOLD state efficiency made by each distinction between the two models. Intermediate complexes and monomeric integrase binding are abbreviated to IC and MIB respectively. In each case an OFF-HOLD operative cycle is simulated, that is,  $a_i = a_x = 10$  followed by  $a_i = a_x = 0.1$ . All remaining non-dimensional parameters are set equal to 1.

cumulatively to observe their influence on HOLD state efficiency. The addition of intermediate complexes is shown to reduce natural re-integration efficiency by  $\sim 11\%$

and thus improves HOLD state efficiency. The addition of monomeric integrase binding is shown to further reduce natural re-integration efficiency by  $\sim 12\%$ . We therefore deduce that the 2:1 integrase:gp3 stoichiometry accounts for the remaining  $\sim 1\%$  that separates the re-integration efficiencies of the two models. With a contribution of  $\sim 1\%$ , the stoichiometry of synaptic complexes follows the previously observed trend of having minimal effect on RAD module dynamics. In contrast, monomeric integrase binding and intermediate DNA:protein complexes provide the vast majority of improvement in HOLD state efficiency for our mechanistic model, in almost equal measure. We note here that the formation of intermediate DNA:protein complexes is, in part, due to the monomeric integrase binding, however, these complexes also arise from our pairwise dimeric integrase binding and monomeric gp3 binding. The disadvantage of minimal intermediate complexes and protein binding pathways is clear with regard to the model of Bonnet et al. [2012]. Following an efficient OFF operation, the concentration of SSRs is restored to a basal level. At this point, there is insufficient protein to hold the system in the  $D_{BP}I_4R_4$  complex and the first interaction that can possibly occur is the dissociation of gp3. This dissociation immediately produces the  $D_{BP}I_4$  complex which is then able to re-integrate. By contrast, the transition from the  $D_{BP}I_4R_2$  complex to the  $D_{BP}I_4$  complex is not as straightforward in our mechanistic model. In the absence of induction, gp3 dissociation produces the intermediate  $D_{BP}I_4R$  complex which itself can potentially give rise to three other complexes, only one of which,  $D_{BP}I_4$ , would then be able to re-integrate.

We ideally require the RAD module to function at 100% efficiency for all three operations. Our results indicate that the efficiency of a HOLD following an OFF switch is proportional to the maintenance of the  $D_{BP}I_4R_2$  complex. This could be problematic when executing a regime whereby increased SSR expression is both induced and ceased in a simultaneous manner. Alternatively, we examine an approach whereby the induction of integrase and gp3 ceases at separate time points. Ceasing induction of gp3 prior to that of integrase is illogical given that gp3 is required to maintain the  $D_{BP}I_4R_2$  complex and prolonged induction of integrase will only facilitate greater re-integration efficiency. Therefore we investigate the effect of ceasing integrase induction prior to that of gp3 on HOLD state efficiency. Figure 3.9 depicts this effect in the form of a plot of natural re-integration efficiency against increasing time intervals,  $\delta t$ , between ceasing integrase induction and gp3 induction. The performance of both the model of Bonnet et al. [2012] and our mechanistic model is plotted, with each y-intercept representing the  $\sim 47\%$  and  $\sim 23\%$  natural re-integration efficiencies for simultaneous cessation ( $\delta t = 0$ ), respectively. Regard-

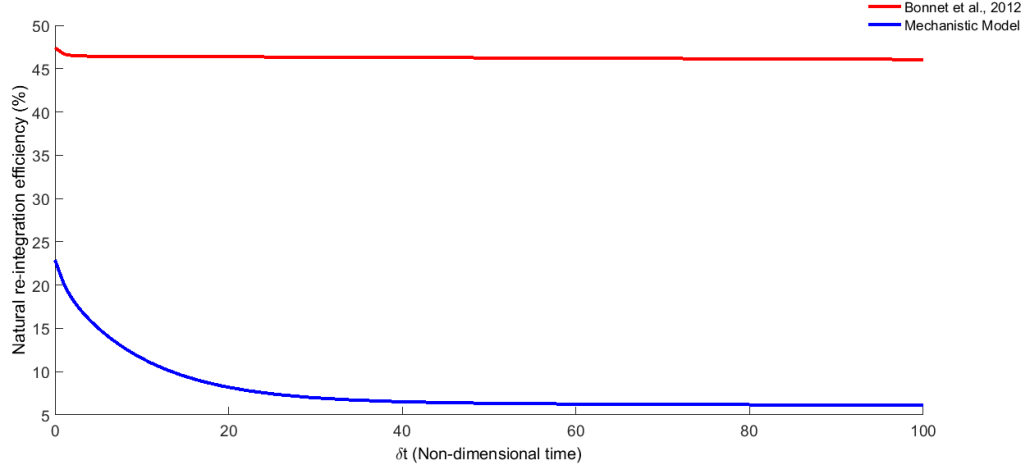


Figure 3.9: RAD module *in vivo* natural re-integration efficiencies for the model of Bonnet et al. [2012] and our mechanistic model. Natural re-integration efficiency is plotted against the time interval between integrase cessation and gp3 cessation,  $\delta t$ . In each case an OFF-HOLD operative cycle is simulated, that is,  $a_i = a_x = 10$  followed by  $a_i = a_x = 0.1$ . All remaining non-dimensional parameters are set equal to 1.

ing the model of Bonnet et al. [2012], as  $\delta t$  increases we observe a minimal reduction in natural re-integration efficiency. This suggests that, although delaying the cessation of gp3 provides a small improvement, any gp3 dissociation that occurs during prolonged induction still provokes an almost immediate re-integration given the inherent transitioning to the  $D_{BP}I_4$  complex. However, in the case of our mechanistic model we observe a significant reduction in natural re-integration efficiency as  $\delta t$  increases. The dissociation of gp3 gives rise to the intermediate  $D_{BP}I_4R$  complex and, with prolonged gp3 induction, the system is therefore weighted in favour of the transition to the  $D_{BP}I_2R$  and  $D_{BP}I_3R$  complexes as well as the reformation of the  $D_{BP}I_4R_2$  complex.

We note that regimes incorporating induction cessation intervals eliminate the HOLD state from the RAD module operative cycle. This may not be desirable for applications regarding biological data storage that are dependent on lasting responses to transient stimuli. However, this may assist in the development of other potential RAD module applications that are not as reliant on HOLD states, such as medical treatments for diseases related to the inheritance of cellular states. We have established that prolonged gp3 induction is capable of reducing re-integration efficiency and thus improving functionality, given that we have considered natural



re-integration to be synonymous with spontaneous switching and ultimately dysfunctionality of the module. However, if we neglect dependency on HOLD states, then harnessing natural re-integration can provide very simple and highly efficient RAD module functionality. Figure 3.10 depicts the dynamical response of the RAD module for 2.5 repeated OFF-ON operative cycles. Here the ON, OFF operations

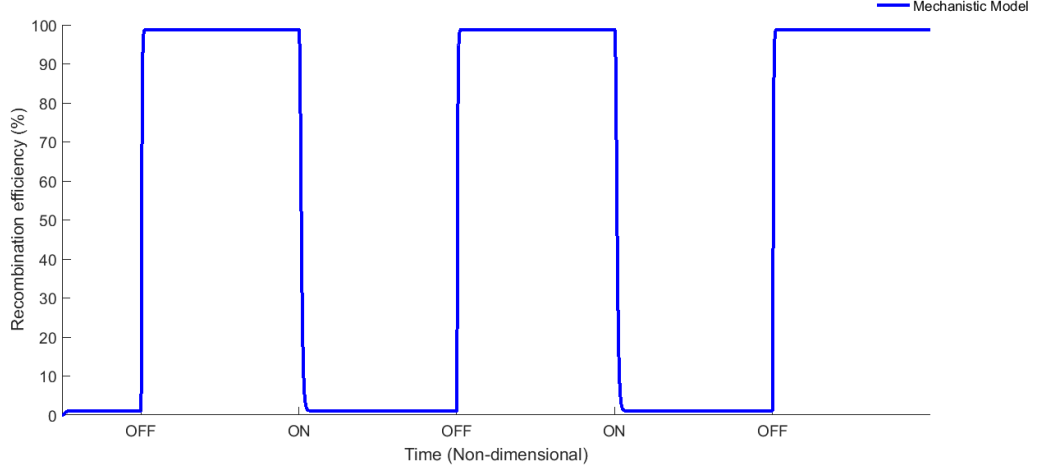


Figure 3.10: RAD module *in vivo* switching efficiencies for our mechanistic model. We simulate an OFF-ON-OFF-ON-OFF operative cycle. For the ON operation  $a_x = 0.1$  and for the OFF operation  $a_x = 10$ . Throughout the operative cycle,  $a_i = 10$ . All remaining non-dimensional parameter values are set equal to 1.

are defined as the cessation of gp3 induction and the induction of gp3 respectively; integrase induction remains constant throughout. Since both operations are mediated, either solely or in part, by integrase, the state of the system is dependent only on gp3 concentration. That is, when there is no induction of gp3 the constant induction of integrase causes a fully efficient ON switch which will remain until gp3 induction causes a fully efficient OFF switch. Induction of gp3 must last for the duration of the desired OFF switch, at which point cessation of gp3 induction is sufficient to cause another fully efficient ON switch through natural re-integration.

### 3.5 Conclusions

We have described the development of a detailed mechanistic model of a rewritable recombinase addressable data module, based on a wide-ranging synthesis of available experimental data on the biomolecular interactions underlying DNA recombination. We demonstrated the capability of this model to match and predict *in vitro* experi-

mental data on recombination efficiencies across a range of different concentrations of integrase and gp3, thus validating its efficacy as a design tool for building future synthetic circuitry. Investigation of *in vivo* recombination dynamics revealed the importance of fully accounting for all mechanistic details in models of DNA recombination, in order to accurately predict the effect of different switching strategies on RAD module performance.

## Chapter 4

# Mechanistic Modelling of a Recombinase-Based Two-Input Temporal Logic Gate

### 4.1 Scientific background

#### 4.1.1 Boolean algebra and logic gates

Binary systems are the language of almost all modern computers and other digital devices [Gillies, 2010]. The origin of such systems can be traced back to ancient Greek philosophers who devised a structured logic paradigm known as propositional logic whereby propositions are classified as TRUE or FALSE. That is, the state of a given proposition is represented in a discrete, binary manner; there are only two possible states that can be achieved. Propositions based on functions of other propositions are constructed via three basic logical connectives, AND, OR and NOT. For example, consider the following statement: *“I will play football this evening if I finish work on time and there are enough other players to play the match”*. Here the proposition of playing football is dependent on the two other propositions regarding finishing work on time and having a sufficient number of players in total. The connective ‘and’ in the statement translates to the logic AND connective by specifying that the outcome of the statement will occur only if both conditional propositions are TRUE. All other combinations of the proposition states will result in the statement being FALSE and hence all possible outcomes can be represented in a table of potential inputs and outputs commonly known as a truth table which, in this case, is representative of the typical AND function (Table 4.1).

In the nineteenth century, the mathematician George Boole formulated a

Finish on time	Enough players	Play football
FALSE	FALSE	FALSE
TRUE	FALSE	FALSE
FALSE	TRUE	FALSE
TRUE	TRUE	TRUE

Table 4.1: Truth table of propositional logic. This proposition is a function of two other propositions which must both be TRUE in order for the overall outcome to be TRUE, thus representing a typical AND logic function.

mathematical framework for the composition and manipulation of logic functions and operations that became known as Boolean algebra. Under this construct, the TRUE and FALSE states are substituted with the numerical values 1 and 0 respectively, thus permitting the application of standard algebraic manipulations from mathematics. The AND and OR binary operators translate to multiplication and addition of Boolean variables respectively and follow the same general associative, commutative and distributive rules of standard algebra. Note that, although the AND function demonstrates intuitive multiplicative results for the Boolean 1 and 0 states, Boolean arithmetic is not generally analogous to standard arithmetic highlighted clearly by the OR function which implies that  $1 + 1 = 1$ . This is a fundamental property of Boolean logic, since no state other than 1 or 0 can be achieved, and therefore should not be viewed as regular numerical addition. For two inputs A and B, the Boolean AND, OR and NOT functions all provide a distinct output O represented by their corresponding truth table (Table 4.2). There exists a total

AND			OR			NOT	
A	B	O	A	B	O	A	O
0	0	0	0	0	0	0	1
1	0	0	1	0	1	1	0
0	1	0	0	1	1		
1	1	1	1	1	1		

Table 4.2: Truth tables for the AND, OR and NOT logic functions. The AND function provides output only when both inputs are present; the OR function provides output whenever at least one input is present; the NOT function provides an output that is the opposite of the given input.

of sixteen Boolean logic functions with each producing distinct functional outputs dependent upon the nature of two specific inputs; logic functions are not limited to two inputs with the number of inputs,  $n$ , giving rise to  $2^n$  outputs. A physical

device that implements Boolean logic functions is referred to as a Boolean logic gate and each of these sixteen distinct gates has a corresponding schematic representation when depicted in electronic circuit designs. Any calculation can be performed through intricate patterns of logic gates and, as a consequence, they form the basis of modern digital computation [Gillies, 2010].

The genetic switches mediated by SSRs are the precursors to synthetic biological logic gates. Both tyrosine and serine recombination mechanisms are capable of eliciting Boolean logic gate functionality within biological cells [Branda and Dymecki, 2004; DuPage et al., 2009; Friedland et al., 2009; Bonnet et al., 2013]. Tyrosine recombinase-based systems are either dependent on host cofactors or are reversible which poses problems with modularity and performance efficiency that have not been identified with serine recombinases. The standard serine recombinase genetic switch is switched on via the induction of integrase which mediates integration and generates a new DNA state. The original state is recovered via the induction of excisionase. By considering integrase as the system input, the standard switch is initially set in its primary state, state 1, in the absence of integrase and then generates the new state, state 2, upon induction (Figure 4.1A). This provides inducible control over two distinct genetic outputs which is particularly valuable in regulating gene expression. There is, however, a maximum of two outputs that can be controlled in this manner given that the system has one integrase input specific to the relevant attachment sites on the DNA strand. As demonstrated, logic gate functions are dependent on two distinct inputs that are either ‘on’ (1) or ‘off’ (0) and therefore biological logic gates require at least two integrase inputs (Figure 4.1B). This requires two pairs of specific attachment sites corresponding to two distinct integrases; multiple identical pairs of attachment sites will be bound by the same integrase and would present the same standard switch mediated by a single integrase input. This also means that combining logic gates to create more sophisticated circuitry can only be achieved using orthogonal gates that employ distinct recombinase inputs [Fernandez-Rodriguez et al., 2015], and hence the functional specifications of characterised recombinases will be crucial in the circuit design process. A recent study has demonstrated that recombinase-based switches that can store 1 b of information, such as the RAD module, are able to provide over 1 B of memory capacity when layered in the appropriate manner. This is made possible by the extensive characterisation of distinct integrases that do not exhibit the crosstalk that may potentially cause the gate to fail, and are therefore completely viable for use as system inputs [Yang et al., 2014].

The precise arrangement and orientation of distinct attachment sites is vital

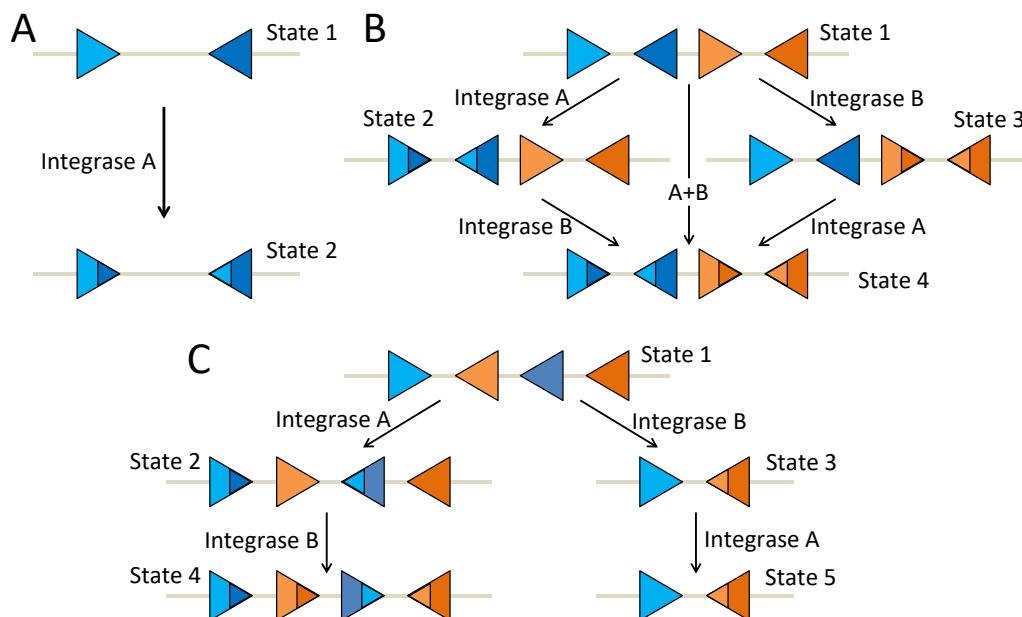


Figure 4.1: Schematic diagrams of one-input and two-input serine integrase-mediated synthetic circuitry. A) the recombinase-based one-input toggle switch. B) the recombinase-based two-input logic gate. C) the recombinase-based two-input temporal logic gate. Blue and orange triangles depict attachment sites corresponding to integrase A and integrase B respectively; grey lines depict DNA strands; arrows depict serine-integrase-mediated integration events.

to realising the desired logic output. By positioning the attachment site pairs sequentially along the DNA strand in an antiparallel manner, the initial sequence of nucleotide bases will encode state 1 in the same way as a standard switch. However, more distinct DNA states can be obtained via induction of each integrase input, integrase A and integrase B. Integrase A will mediate the inversion of the genetic sequence between its corresponding attB and attP sites and produce the encoded state 2; integrase B will mediate the equivalent reaction to produce state 3. Simultaneous induction of integrase A and integrase B results in the inversion of both corresponding genetic sequences which encodes the final output, state 4. The same four distinct DNA states are achieved if the timing of the induction events is staggered, rather than simultaneous, since state 1 and state 2 both reach state 4 in the presence of the appropriate integrase. This dynamical behaviour defines a truth table for the standard two-input biological logic gate (Table 4.3A). State transitioning is a unidirectional process unless excisionase-mediated reset functions are incorporated, however this will invariably compromise the dynamical behaviour of the logic gate or toggle switch in question [Bonnet et al., 2012; Bowyer et al.,

A			B		
A	B	O	A	B	O
0	0	State 1	0	0	State 1
1	0	State 2	1	0	State 2
0	1	State 3	0	1	State 3
1	1	State 4	1 <sup>*</sup>	1	State 4
			1	1 <sup>*</sup>	State 5

Table 4.3: Truth tables for the standard and temporal biological logic gates. A) The standard logic gate provides the typical four outputs as DNA states that could be engineered to deliver any logic function. B) The temporal logic gate also provides four distinct outputs since at least two of the states are identical; stars denote which of the two integrase inputs is induced first.

2016].

The most interesting distinction to be made between the biological logic gate and the idealised Boolean logic gate that it seeks to emulate is that the biological logic gate is able to encode up to four distinct states as opposed to a maximum of two states. That is, although the integrase inputs driving the biological gate occupy binary states, the outputs are four distinct genetic sequences each of which could potentially code for a different molecular product if engineered in the appropriate manner. Indeed, designing genetic sequences and attachment site pairs that produce desired gene expression only when the configuration reaches that of state 4 would encompass the AND gate function (i.e. state 1 = 0, state 2 = 0, state 3 = 0, state 4 = 1). This might be particularly useful in cases where extra stringency on the control of gene expression is required. On the other hand, restricting the scope of biological output to one of two binary states may be viewed as prosaic in light of the potential variety of operations.

The biological logic gate has the capacity for even further functional advantages when considering the temporal induction of integrase inputs [Hsiao et al., 2016]. As previously described, sequential positioning of two distinct attachment site pairs will provide a maximum of four outputs due to staggered induction events giving rise to the same end state, but alternative initial arrangements are capable of providing additional information. For example, overlapping attachment sites has the ability to define integration pathways culminating in two distinct end states, given the appropriate initial orientation of the attachment sites (Figure 4.1C). Such circuit designs exploit the DNA inversion event by placing attachment sites in the intermediate genetic sequence between two corresponding attachment sites thus resulting in an inverted genetic sequence giving rise to a new DNA state as well as an

inverted attachment site whose role in any subsequent integration events is altered and may potentially partake in either inversion or deletion.

Positioning the attachment site pairs in an overlapping arrangement on the DNA strand with one antiparallel pair (corresponding to integrase A) and one parallel pair (corresponding to integrase B) presents a similar state 1 to that of the standard logic gate design. However, in this case integrase A will mediate the inversion of the genetic sequence between its corresponding attB and attP sites and produce the encoded state 2 comprised of both pairs of attachment sites in an antiparallel arrangement. Integrase B will, instead, mediate the deletion of the genetic sequence between its corresponding attachment sites to produce state 3, comprised of two disparate attachment sites. Induction of integrase B following integrase A will then transition state 2 to state 4 via a secondary inversion event due to the new antiparallel orientation of the corresponding attachment sites. Integrase A is unable to perform integration following induction of integrase B since there is no longer an appropriate pair of attachment sites to target and thus state 5 is identical to state 3. Hence, we have a temporal logic gate capable of delivering five outputs in response to two inputs (Table 4.3B) however, this functionality is restricted to at least two of the five outputs being identical which, biologically, presents the potential for inducible control over four molecular products, as is the case for the standard logic gate. Thus, although the number of distinct outputs has not improved, the temporal logic gate is able to infer both the order of induction events and the time between each event since staggered induction does not produce identical end states (Figure 4.2). A temporal logic gate therefore has unique functional properties that make it highly suitable for synthetic biosensor applications.

An important factor to consider with respect to temporal logic gates is the time interval between induction events since simultaneous induction of input integrases will likely result in a split end state whereby both state 4 and state 5 are accessed. Ideally, the appropriate induction will facilitate maximal transitioning to the desired DNA state however, the conditions required to achieve this are not obvious. Mathematical models can examine large arrays of performance criteria in order to determine optimal inputs and provide operational profiles that can inform the selection of the most suitable circuit designs based on the desired outputs.

#### 4.1.2 Existing logic gate systems

Most approaches to engineering cellular logic employ two-input circuit designs in which the orientation and arrangement of specific attachment sites, analogous to gate-gate layering in conventional electronics, is used to mediate all conceivable



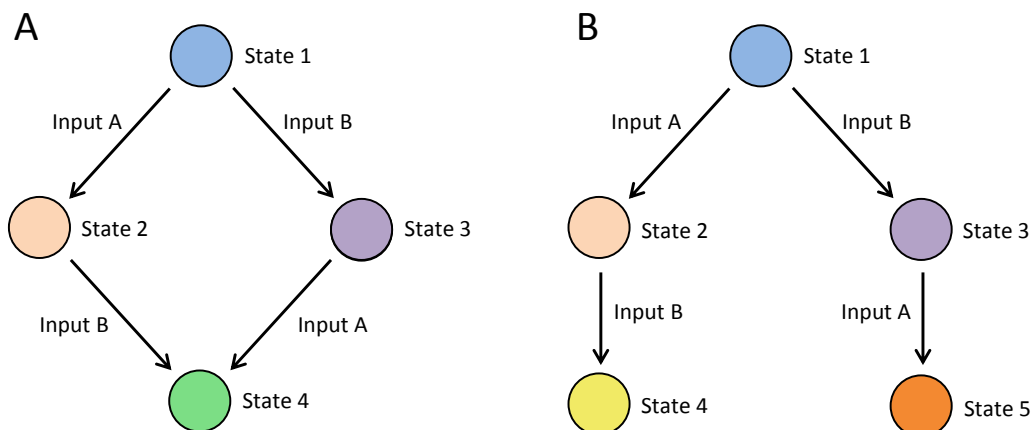


Figure 4.2: Schematic diagrams summarising biological logic outputs. A) the standard logic gate transitions through four distinct states with the end state accessed regardless of the timing of integrase inputs. B) the temporal logic gate transitions through five states with certain states only accessible via specific induction time schedules.

logic functions [Tamsir et al., 2011; Moon et al., 2012]. Transcriptional systems have facilitated the construction of numerous logic functions including NOT-IF, NOT-IF-IF, NAND, OR, NOR and INVERTER gates by virtue of synthetic transcription factors in mammalian cells [Kramer et al., 2004a]. A similar approach based on chimeric transcription factors can enable simple transcriptional AND gating [Shis et al., 2014]. Mammalian logic circuits can also exploit protein splicing mediated by self-splicing protein domains known as inteins to build AND gates [Lohmueller et al., 2012] and has been extended to realise a 3-input AND gate [Lienert et al., 2014]. Other transcriptional AND gates are achieved through the combinatorial presence of both a T7 polymerase and a suppressor tRNA, *supD* [Anderson et al., 2006] as well as through promoter inputs controlling *hrpR* and *hrpS* in conjunction with an *hrpL* promoter output [Wang et al., 2014]. The first example of a genetic 2-4 decoder expressing fluorescent protein promoter outputs has also been demonstrated in eukaryotic cells [Guinn and Bleris, 2014]. Several alternative mechanisms for implementing biological logic have been devised that are mediated by ribosomes [Rackham and Chin, 2005], RNA [Win and Smolke, 2008; Benenson et al., 2004], RNAi [Rinaudo et al., 2007; Xie et al., 2010, 2011] and deoxyribozymes [Stojanovic and Stefanovic, 2003].

Scaling up the complexity of electronic circuitry involves layered combinations of logic gates which can be replicated biologically to create advanced circuitry

capable of adding and subtracting digital information. This has been demonstrated through a combination of transcriptional and translational control whereby two N-IMPLY gates were layered to create an XOR circuit which was then developed into a half-adder circuit by virtue of an additional AND gate [Auslender and Fussenegger, 2013]. Transcriptional AND gates implementing two promoter inputs can be layered, with promoter outputs providing inputs to downstream gates, to achieve all logic functions. Combinations of AND gates permitted the implementation of 3- and 4-input AND gates with the 4-input AND representing the largest synthetic gene circuit constructed at the time, in terms of the number of regulatory proteins used [Moon et al., 2012]. The advent of greater system complexity requires greater consideration of parts modularity. One potential solution could be wiring logic circuits across cellular populations using quorum sensing circuitry to overcome such issues by installing different system components in different cells, rather than complete assembly within individual cells, and employing cell-cell communication strategies to realise overall logic functions [Tabor et al., 2009; Regot et al., 2011; Tamsir et al., 2011].

Many of the aforementioned transcriptional logic circuits are transient in their responses and are generally unable to provide lasting outputs in the absence of induction. However, integration of memory and logic has been demonstrated through a ‘push-on push-off’ switch that connects a bistable memory module with a transcriptional NOR gate module. The circuit is able to alternate between green and red fluorescent protein outputs in response to sequential induction via pulses of ultraviolet light and can hold the current state for the duration of the interval between pulses. That said, just three consecutive switching events were achieved experimentally and it was observed that efficiency decreased with each event [Lou et al., 2010]. As a result, DNA-based systems may be preferable for robust biological computation coupled with cellular memory as demonstrated by DNA recombination circuits such as the RAD module [Bonnet et al., 2012].

Although it remains to verify that SSRs can operate with the same reliability as bistable transcriptional circuits, given the practical difficulties associated with efficient integrase-excisionse-mediated reset functions [Bonnet et al., 2012; Bowyer et al., 2016], the overwhelming advantage of exploiting the stability of DNA suggests that SSRs have the edge when it comes to engineering biological logic gates [Friedland et al., 2009; Purcell and Lu, 2014]. The unidirectional recombinases Bxb1 and  $\phi$ C31 mediate control over promoters, terminators and other transcriptional gene regulatory elements to perform all sixteen two-input logic gate functions. For example, an AND gate can be constructed by placing one inverted promoter and

one terminator in tandem upstream of an output gene, each flanked by distinct attachment site pairs corresponding to Bxb1 and  $\phi$ C31. Expression of the SSR corresponding to the sites flanking the inverted promoter will cause inversion and hence establish regular promoter orientation however, the action of the promoter will be blocked by the terminator. Expression of the SSR corresponding to the sites flanking the terminator will also cause inversion and hence the terminator will no longer block the upstream promoter however, the promoter will remain inverted and the gene of interest is not expressed. When both Bxb1 and  $\phi$ C31 are expressed simultaneously, both the inverted promoter and the terminator will undergo inversion resulting in unrestricted transcription of the gene of interest by the upstream promoter and hence the target gene will only be transcribed in the presence of both inputs.

Two inverted promoters placed in tandem and controlled in the same fashion will produce an OR gate function and thus all logic gates can be formed by virtue of a specific initial arrangement of attachment sites and transcriptional elements. Since the inversion events rewrite the genetic sequence, the state of the system is stored stably within the DNA and permits sequential memory logic that will give, in the case of the AND gate, the desired output if the two inputs were ever expressed as opposed to the equivalent memory-less circuits which provides the desired output if the two inputs are expressed at a given time [Siuti et al., 2013]. A similar approach is capable of achieving the same variety of logic gate functions through inversion of terminators only. By positioning a promoter upstream of two terminators, flanked by distinct attachment site pairs, and a target gene, the gene will be expressed only when both SSRs have been induced and cause inversion of their respective terminator hence capturing the same AND gate dynamics described previously [Bonnet et al., 2013].

These recombinase-based systems are able to match the reliability of bistable transcriptional systems since the inversion events that drive the logic gates are mediated solely by the selected integrases. Integrase alone is sufficient to mediate integration reactions and hence integrase-mediated inversions, insertions and deletions are dependent purely on the concentration of integrase which, assuming the appropriate induction is delivered, can give rise to maximal efficiency and, in turn, reliability. Therefore, reliability issues arise with greater significance when attempting to incorporate excisionase-mediated reset functions [Bonnet et al., 2012; Bowyer et al., 2016]. Reset functions are useful however, as they, theoretically, allow the system to re-establish previously accessed states and thus offer a chain of states that can be transitioned through in any consecutive order as opposed to sequential memory

logic gates that transition in a forward, unidirectional manner. Until the dynamical nuances of integrase-excisionse multiplexed systems are characterised, circuit designs are likely to benefit from sole focus towards integrase-mediated recombination and the exclusion of any excisionase-mediate reset functionality.

#### 4.1.3 Existing logic gate models

Temporal logic gates generate as many distinct outputs as standard logic gates, but are also capable of inferring additional information due to the significance of the timing of input induction events. Both the temporal order of induction events and the length of time allowed between them can influence the output of the system. The first experimentally validated temporal logic gate is capable of recording analogue signal timing and the duration of sequential induction events with respect to a bacterial cell population [Hsiao et al., 2016].

The temporal logic gate demonstrates ‘A then B’ logic in *E. coli* through a system of two orthogonal integrases (integrase A, TP901-1, and B, Bxb1). Being a temporal logic gate, the system is capable of accessing five genetic states (Figure 4.2B) however, the circuit design comprises just two states of interest in order to elicit reliable timing and recording of induction events. The initial state is comprised of one inverted terminator and one inverted promoter flanked by the attB and attP sites corresponding to integrases A and B which overlap one another and are themselves flanked by the genes coding the mKate-2 RFP and superfolder-GFP proteins, neither of which are expressed due to the inverted terminator blocking transcription of RFP via the inverted promoter. The antiparallel orientation of the integrase A attachment sites is selected in order to permit inversion whereas the parallel orientation of the integrase B attachment sites is selected in order to permit deletion. Therefore, induction of integrase B deletes both the inverted promoter and the integrase A attP site from the DNA strand, leaving the inverted terminator behind in a new genetic state,  $S_b$  (state 3), which again results in no fluorescent protein expression. Subsequent integrase A induction elicits no further response since integrase A has only the remaining attB site on the DNA strand to target. However, initial induction of integrase A mediates inversion of the genetic sequence between the corresponding attachment sites containing the inverted terminator and the integrase B attB site. As a result, the terminator is unable to block transcription via the inverted promoter and RFP is expressed,  $S_a$  (state 2). Subsequent integrase B induction is then able to mediate a secondary inversion event, since the associated attB site was inverted along with the terminator via integrase A, which results in inversion of the terminator-promoter pair and triggers expression of GFP,  $S_{ab}$  (state

4). Table 4.4 gives the truth table for this particular temporal logic gate. Switching

A	B	O
0	0	State 1: 0
1	0	State 2: RFP
0	1	State 3: 0
1*	1	State 4: GFP
1	1*	State 5: 0

Table 4.4: Truth table for the synthetic temporal logic gate reported in [Hsiao et al., 2015]. The output of the logic gate is dependent on the order of induction events and the time delay between events, with the star notation denoting the input induced first.

between four distinct states is achievable via standard logic gates however, such gates produce the same state with simultaneous input induction and regardless of delays between induction events. That is, the benefits of this particular temporal system allow the user to infer which induction event has occurred first, since ‘A then B’ operations give GFP output as opposed to no output for ‘B then A’ operations, and also the time delay between induction events, since the ratio of RFP to GFP output will vary for varying time delays. A simple stochastic model of this circuit was created to help develop better intuition for overall circuit behaviour, but this model does not account for any specific molecular interactions between the integrases and the DNA, instead representing integrase activity as probabilities based on concentration [Hsiao et al., 2016]. This model was shown to be effective for predicting overall final population fractions as well as the forward experimental design of the system. However, the inherent limitations of the model design mean that the circuit cannot be simulated on a molecular scale, and timescales with regards to specific molecular interactions cannot be incorporated. In taking a more mechanistic modelling approach, we model the two-integrase temporal logic gate circuit developed in Hsiao et al. [2016] by integrating multiple DNA recombination interactions from our validated mechanistic model of the RAD module [Bowyer et al., 2015, 2016]. We demonstrate that the mechanistic model successfully captures key dynamical features of circuit time course trajectories derived from *in vivo* experimental data, thus improving our capability to perform model-aided integrase circuit design.

## 4.2 Formulating an *in vivo* reaction network of a recombinase-based temporal logic gate

The model of [Hsiao et al., 2016] describes the transitioning of the temporal logic gate from the original DNA state ( $S_0$ ) to each of the three end states ( $S_a$ ,  $S_b$  and  $S_{ab}$ ) via three corresponding rate constants describing inversion mediated by TP901-1 (integrase A), and both deletion and inversion mediated by Bxb1 (integrase B). We replace these all-encompassing parameters with a mechanistic integration reaction structure, in which inversion and deletion events are initiated by the binding of one integrase dimer at each of the associated attachment sites (four integrase monomers in total) and are strictly unidirectional [Groth and Calos, 2004; Olorunniji et al., 2012] (Fig. 4.3). We account for dimerisation of both monomeric SSRs, allowing

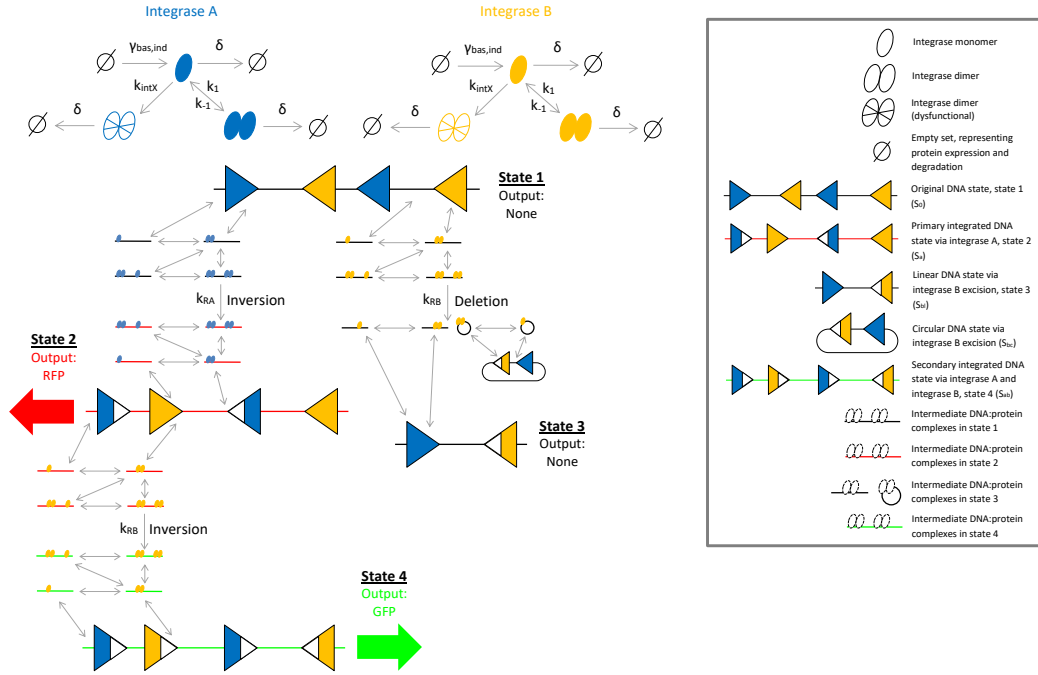


Figure 4.3: Schematic diagram of the mechanistic two-input logic gate reaction network. The sequence of DNA:protein interactions facilitating inversion and deletion, taken from Bowyer et al. [2015], enables the system to transition from the original DNA state (state 1) to the three genetically-differentiated DNA states (state 2, state 3, state 4). Expression and degradation of integrase A and B are denoted by  $\gamma_{bas,ind}$  and  $\delta$  respectively. Intermediate DNA:protein complexes with red and green DNA strands are associated with state 2 and state 4 respectively. Up to four integrase monomers are able to bind to free DNA, depicted by dashed outlines in the legend. Single and double grey arrows depict irreversible and reversible reactions respectively. Figure adapted from [Hsiao et al., 2016].

for both monomeric and dimeric integrase binding to DNA attachment sites. This process is widely supported in the experimental literature on DNA recombination [Fogg et al., 2014; Ghosh et al., 2005, 2008; Khaleel et al., 2011]. Thus, four distinct intermediate DNA:protein complexes can potentially be formed in facilitating recombination via this combination of monomeric and dimeric integrase binding. These complexes are depicted by the smaller DNA strands in Fig. 4.3, with up to four integrase monomers bound. We also include the formation of a dysfunctional dimer by both integrases, which is subject to the same degradation as its functional counterpart; this was shown to significantly improve the fit of model simulations to experimental data in our previous study [Bowyer et al., 2015, 2016]. Deletion gives rise to two distinct genetic products, the remaining linear sequence of the target DNA and the excised circular DNA loop, and hence we account for two disparate  $S_b$  states,  $S_{bl}$  (linear) and  $S_{bc}$  (circular). We refer to the DNA states  $S_0$ ,  $S_a$ ,  $S_{bl}$  and  $S_{ab}$  as state 1, state 2, state 3 and state 4 respectively, noting that this system is void of a fifth DNA state due to the ‘dead end’ state, state 3, which is unable to transition to any subsequent state. We do not account for  $S_{bc}$  as a system state since it is separate from the original DNA sequence that is intended to be manipulated.

Model validation against experimental data for an *in vivo* recombinase-based system presents a number of factors that require careful consideration. Cellular recombination *in vivo* is dependent on the expression and degradation of recombinase proteins over time, thus contributing two additional parameters ( $\gamma$ ,  $\delta$ ) to the model. In the absence of induction, *in vivo* system exhibit background expression of SSRs, commonly resulting in basal system output. Thus we include model parameters describing both basal ( $\gamma_{bas}$ ) and induced ( $\gamma_{ind}$ ) expression of the two integrases, allowing for non-zero model output when simulating experiments void of integrase induction.

### 4.3 Constructing a mechanistic model of the temporal logic gate

Our mechanistic model is constructed through the application of mass action kinetics to the following biochemical equations arising from the reaction network in Fig. 4.3:



$$I_{2A} \xrightarrow{\delta} \emptyset, \quad (4.3)$$

$$I_A + I_A \xrightarrow{k_{\text{int}X}} I_{2AX}, \quad (4.4)$$

$$I_{2AX} \xrightarrow{\delta} \emptyset, \quad (4.5)$$

$$I_B \xrightleftharpoons[\gamma_{\text{bas,ind}}]{\delta} \emptyset, \quad (4.6)$$

$$I_B + I_B \xrightleftharpoons[k_{-1}]{k_1} I_{2B}, \quad (4.7)$$

$$I_{2B} \xrightarrow{\delta} \emptyset, \quad (4.8)$$

$$I_B + I_B \xrightarrow{k_{\text{int}X}} I_{2BX}, \quad (4.9)$$

$$I_{2BX} \xrightarrow{\delta} \emptyset, \quad (4.10)$$

$$S_0 + I_{2A} \xrightleftharpoons[k_{-4}]{k_4} S_0 I_{2A}, \quad (4.11)$$

$$S_0 I_{2A} + I_{2A} \xrightleftharpoons[k_{-5}]{k_5} S_0 I_{4A}, \quad (4.12)$$

$$S_0 + I_A \xrightleftharpoons[k_{-13}]{k_{13}} S_0 I_A, \quad (4.13)$$

$$S_0 I_A + I_A \xrightleftharpoons[k_{-14}]{k_{14}} S_0 I_{2A}, \quad (4.14)$$

$$S_0 I_{2A} + I_A \xrightleftharpoons[k_{-15}]{k_{15}} S_0 I_{3A}, \quad (4.15)$$

$$S_0 I_{3A} + I_A \xrightleftharpoons[k_{-16}]{k_{16}} S_0 I_{4A}, \quad (4.16)$$

$$S_0 I_{4A} \xrightarrow{k_{RA}} S_a I_{4A}, \quad (4.17)$$

$$S_a + I_{2A} \xrightleftharpoons[k_{-4}]{k_4} S_a I_{2A}, \quad (4.18)$$

$$S_a I_{2A} + I_{2A} \xrightleftharpoons[k_{-5}]{k_5} S_a I_{4A}, \quad (4.19)$$

$$S_a + I_A \xrightleftharpoons[k_{-13}]{k_{13}} S_a I_A, \quad (4.20)$$

$$S_a I_A + I_A \xrightleftharpoons[k_{-14}]{k_{14}} S_a I_{2A}, \quad (4.21)$$

$$S_a I_{2A} + I_A \xrightleftharpoons[k_{-15}]{k_{15}} S_a I_{3A}, \quad (4.22)$$

$$S_a I_{3A} + I_A \xrightleftharpoons[k_{-16}]{k_{16}} S_a I_{4A}, \quad (4.23)$$

$$S_0 + I_{2B} \xrightleftharpoons[k_{-20}]{k_{20}} S_0 I_{2B}, \quad (4.24)$$



$$S_0 I_{2B} + I_{2B} \xrightleftharpoons[k_{-21}]{k_{21}} S_0 I_{4B}, \quad (4.25)$$

$$S_0 + I_B \xrightleftharpoons[k_{-22}]{k_{22}} S_0 I_B, \quad (4.26)$$

$$S_0 I_B + I_B \xrightleftharpoons[k_{-23}]{k_{23}} S_0 I_{2B}, \quad (4.27)$$

$$S_0 I_{2B} + I_B \xrightleftharpoons[k_{-24}]{k_{24}} S_0 I_{3B}, \quad (4.28)$$

$$S_0 I_{3B} + I_B \xrightleftharpoons[k_{-25}]{k_{25}} S_0 I_{4B}, \quad (4.29)$$

$$S_0 I_{4B} \xrightarrow{k_{RB}} S_{bl} I_{2B} + S_{bc} I_{2B}, \quad (4.30)$$

$$S_{bl} + I_{2B} \xrightleftharpoons[k_{-20}]{k_{20}} S_{bl} I_{2B}, \quad (4.31)$$

$$S_{bl} + I_B \xrightleftharpoons[k_{-22}]{k_{22}} S_{bl} I_B, \quad (4.32)$$

$$S_{bl} I_B + I_B \xrightleftharpoons[k_{-23}]{k_{23}} S_{bl} I_{2B}, \quad (4.33)$$

$$S_{bc} + I_{2B} \xrightleftharpoons[k_{-20}]{k_{20}} S_{bc} I_{2B}, \quad (4.34)$$

$$S_{bc} + I_B \xrightleftharpoons[k_{-22}]{k_{22}} S_{bc} I_B, \quad (4.35)$$

$$S_{bc} I_B + I_B \xrightleftharpoons[k_{-23}]{k_{23}} S_{bc} I_{2B}, \quad (4.36)$$

$$S_a + I_{2B} \xrightleftharpoons[k_{-20}]{k_{20}} S_a I_{2B}, \quad (4.37)$$

$$S_a I_{2B} + I_{2B} \xrightleftharpoons[k_{-21}]{k_{21}} S_a I_{4B}, \quad (4.38)$$

$$S_a + I_B \xrightleftharpoons[k_{-22}]{k_{22}} S_a I_B, \quad (4.39)$$

$$S_a I_B + I_B \xrightleftharpoons[k_{-23}]{k_{23}} S_a I_{2B}, \quad (4.40)$$

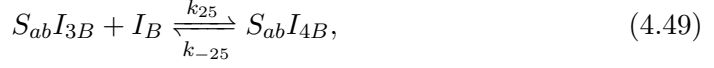
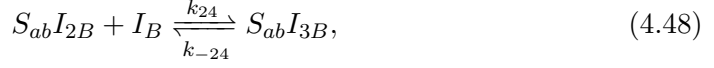
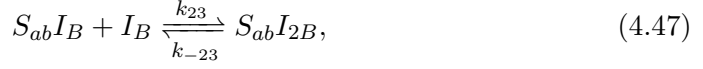
$$S_a I_{2B} + I_B \xrightleftharpoons[k_{-24}]{k_{24}} S_a I_{3B}, \quad (4.41)$$

$$S_a I_{3B} + I_B \xrightleftharpoons[k_{-25}]{k_{25}} S_a I_{4B}, \quad (4.42)$$

$$S_a I_{4B} \xrightarrow{k_{RB}} S_{ab} I_{4B}, \quad (4.43)$$

$$S_{ab} + I_{2B} \xrightleftharpoons[k_{-20}]{k_{20}} S_{ab} I_{2B}, \quad (4.44)$$

$$S_{ab} I_{2B} + I_{2B} \xrightleftharpoons[k_{-21}]{k_{21}} S_{ab} I_{4B}, \quad (4.45)$$



where reaction rates are denoted by the corresponding numbered  $k$ , which are retained from [Bowyer et al., 2015, 2016] where necessary;  $S$  denotes DNA with subscripts corresponding to the four distinct states (0, a, bl, ab);  $I$  denotes integrase with subscripts corresponding to the number of monomers (1, 2, 3, 4), TP901-1 or Bxb1 (A or B) and/or dysfunctionality (X). This produces the following system of 35 ODEs that are solved numerically to provide a deterministic model output:

$$\begin{aligned} \frac{d[I_A]}{dt} = & \gamma_{\text{bas,ind}} - \delta[I_A] + 2k_{-1}[I_{2A}] - 2k_1[I_A]^2 - k_{\text{intX}}[I_A]^2 + \\ & k_{-13}[S_0I_A] - k_{13}[S_0][I_A] + k_{-13}[S_aI_A] - k_{13}[S_a][I_A] + \\ & k_{-14}[S_0I_{2A}] - k_{14}[S_0I_A][I_A] + k_{-14}[S_aI_{2A}] - k_{14}[S_aI_A][I_A] + \\ & k_{-15}[S_0I_{3A}] - k_{15}[S_0I_{2A}][I_A] + k_{-15}[S_aI_{3A}] - k_{15}[S_aI_{2A}][I_A] + \\ & k_{-16}[S_0I_{4A}] - k_{16}[S_0I_{3A}][I_A] + k_{-16}[S_aI_{4A}] - k_{16}[S_aI_{3A}][I_A], \end{aligned} \quad (4.50)$$

$$\begin{aligned} \frac{d[I_{2A}]}{dt} = & k_1[I_A]^2 - k_{-1}[I_{2A}] - \delta[I_{2A}] + \\ & k_{-4}[S_0I_{2A}] - k_4[S_0][I_{2A}] + k_{-4}[S_aI_{2A}] - k_4[S_a][I_{2A}] + \\ & k_{-5}[S_0I_{4A}] - k_5[S_0I_{2A}][I_{2A}] + k_{-5}[S_aI_{4A}] - k_5[S_aI_{2A}][I_{2A}], \end{aligned} \quad (4.51)$$

$$\begin{aligned} \frac{d[S_0]}{dt} = & k_{-4}[S_0I_{2A}] - k_4[S_0][I_{2A}] + k_{-13}[S_0I_A] - k_{13}[S_0][I_A] + \\ & k_{-20}[S_0I_{2B}] - k_{20}[S_0][I_{2B}] + k_{-22}[S_0I_B] - k_{22}[S_0][I_B], \end{aligned} \quad (4.52)$$

$$\begin{aligned} \frac{d[S_0I_{2A}]}{dt} = & k_4[S_0][I_{2A}] - k_{-4}[S_0I_{2A}] + k_{-5}[S_0I_{4A}] - k_5[S_0I_{2A}][I_{2A}] + \\ & k_{14}[S_0I_A][I_A] - k_{-14}[S_0I_{2A}] + k_{-15}[S_0I_{3A}] - k_{15}[S_0I_{2A}][I_A], \end{aligned} \quad (4.53)$$

$$\frac{d[S_0I_{4A}]}{dt} = k_5[S_0I_{2A}][I_{2A}] - k_{-5}[S_0I_{4A}] + k_{16}[S_0I_{3A}][I_A] - k_{-16}[S_0I_{4A}] - k_{RA}[S_0I_{4A}], \quad (4.54)$$

$$\begin{aligned} \frac{d[S_a]}{dt} = & k_{-4}[S_aI_{2A}] - k_4[S_a][I_{2A}] + k_{-13}[S_aI_A] - k_{13}[S_a][I_A] + \\ & k_{-20}[S_aI_{2B}] - k_{20}[S_a][I_{2B}] + k_{-22}[S_aI_B] - k_{22}[S_a][I_B], \end{aligned} \quad (4.55)$$

$$\begin{aligned} \frac{d[S_aI_{2A}]}{dt} = & k_4[S_a][I_{2A}] - k_{-4}[S_aI_{2A}] + k_{-5}[S_aI_{4A}] - k_5[S_aI_{2A}][I_{2A}] + \\ & k_{14}[S_aI_A][I_A] - k_{-14}[S_aI_{2A}] + k_{-15}[S_aI_{3A}] - k_{15}[S_aI_{2A}][I_A], \end{aligned} \quad (4.56)$$

$$\frac{d[S_aI_{4A}]}{dt} = k_5[S_aI_{2A}][I_{2A}] - k_{-5}[S_aI_{4A}] + k_{16}[S_aI_{3A}][I_A] - k_{-16}[S_aI_{4A}] + k_{RA}[S_0I_{4A}], \quad (4.57)$$

$$\frac{d[I_{2AX}]}{dt} = k_{\text{intX}}[I_A]^2 - \delta[I_{2AX}], \quad (4.58)$$

$$\frac{d[S_0I_A]}{dt} = k_{13}[S_0][I_A] - k_{-13}[S_0I_A] - k_{14}[S_0I_A][I_A] + k_{-14}[S_0I_{2A}], \quad (4.59)$$

$$\frac{d[S_0 I_{3A}]}{dt} = k_{15}[S_0 I_{2A}][I_A] - k_{-15}[S_0 I_{3A}] - k_{16}[S_0 I_{3A}][I_A] + k_{-16}[S_0 I_{4A}], \quad (4.60)$$

$$\frac{d[S_a I_A]}{dt} = k_{13}[S_a][I_A] - k_{-13}[S_a I_A] - k_{14}[S_a I_A][I_A] + k_{-14}[S_a I_{2A}], \quad (4.61)$$

$$\frac{d[S_a I_{3A}]}{dt} = k_{15}[S_a I_{2A}][I_A] - k_{-15}[S_a I_{3A}] - k_{16}[S_a I_{3A}][I_A] + k_{-16}[S_a I_{4A}], \quad (4.62)$$

$$\begin{aligned} \frac{d[I_B]}{dt} = & \gamma_{\text{bas,ind}} - \delta[I_B] + 2k_{-1}[I_{2B}] - 2k_1[I_B]^2 - k_{\text{intX}}[I_B]^2 + \\ & k_{-22}[S_0 I_B] - k_{22}[S_0][I_B] + k_{-22}[S_a I_B] - k_{22}[S_a][I_B] + \\ & k_{-23}[S_0 I_{2B}] - k_{23}[S_0 I_B][I_B] + k_{-23}[S_a I_{2B}] - k_{23}[S_a I_B][I_B] + \\ & k_{-24}[S_0 I_{3B}] - k_{24}[S_0 I_{2B}][I_B] + k_{-24}[S_a I_{3B}] - k_{24}[S_a I_{2B}][I_B] + \\ & k_{-25}[S_0 I_{4B}] - k_{25}[S_0 I_{3B}][I_B] + k_{-25}[S_a I_{4B}] - k_{25}[S_a I_{3B}][I_B] + \\ & k_{-22}[S_b I_B] - k_{22}[S_b][I_B] + k_{-22}[S_{ab} I_B] - k_{22}[S_{ab}][I_B] + \\ & k_{-23}[S_b I_{2B}] - k_{23}[S_b I_B][I_B] + k_{-23}[S_{ab} I_{2B}] - k_{23}[S_{ab} I_B][I_B] + \\ & k_{-24}[S_b I_{3B}] - k_{24}[S_b I_{2B}][I_B] + k_{-24}[S_{ab} I_{3B}] - k_{24}[S_{ab} I_{2B}][I_B] + \\ & k_{-25}[S_b I_{4B}] - k_{25}[S_b I_{3B}][I_B] + k_{-25}[S_{ab} I_{4B}] - k_{25}[S_{ab} I_{3B}][I_B], \end{aligned} \quad (4.63)$$

$$\begin{aligned} \frac{d[I_{2B}]}{dt} = & k_{19}[I_B]^2 - k_{-19}[I_{2B}] - \delta[I_{2B}] + \\ & k_{-20}[S_0 I_{2B}] - k_{20}[S_0][I_{2B}] + k_{-20}[S_a I_{2B}] - k_{20}[S_a][I_{2B}] + \\ & k_{-21}[S_0 I_{4B}] - k_{21}[S_0 I_{2B}][I_{2B}] + k_{-21}[S_a I_{4B}] - k_{21}[S_a I_{2B}][I_{2B}] + \\ & k_{-20}[S_b I_{2B}] - k_{20}[S_b][I_{2B}] + k_{-20}[S_{ab} I_{2B}] - k_{20}[S_{ab}][I_{2B}] + \\ & k_{-21}[S_b I_{4B}] - k_{21}[S_b I_{2B}][I_{2B}] + k_{-21}[S_{ab} I_{4B}] - k_{21}[S_{ab} I_{2B}][I_{2B}], \end{aligned} \quad (4.64)$$

$$\begin{aligned} \frac{d[S_0 I_{2B}]}{dt} = & k_{20}[S_0][I_{2B}] - k_{-20}[S_0 I_{2B}] + k_{-21}[S_0 I_{4B}] - k_{21}[S_0 I_{2B}][I_{2B}] + \\ & k_{23}[S_0 I_B][I_B] - k_{-23}[S_0 I_{2B}] + k_{-24}[S_0 I_{3B}] - k_{24}[S_0 I_{2B}][I_B], \end{aligned} \quad (4.65)$$

$$\frac{d[S_0 I_{4B}]}{dt} = k_{21}[S_0 I_{2B}][I_{2B}] - k_{-21}[S_0 I_{4B}] + k_{25}[S_0 I_{3B}][I_B] - k_{-25}[S_0 I_{4B}] - k_{RB}[S_0 I_{4B}], \quad (4.66)$$

$$\frac{d[S_b]}{dt} = k_{-20}[S_b I_{2B}] - k_{20}[S_b][I_{2B}] + k_{-22}[S_b I_B] - k_{22}[S_b][I_B], \quad (4.67)$$

$$\begin{aligned} \frac{d[S_b I_{2B}]}{dt} = & k_{20}[S_b][I_{2B}] - k_{-20}[S_b I_{2B}] + k_{-21}[S_b I_{4B}] - k_{21}[S_b I_{2B}][I_{2B}] + \\ & k_{23}[S_b I_B][I_B] - k_{-23}[S_b I_{2B}] + k_{-24}[S_b I_{3B}] - k_{24}[S_b I_{2B}][I_B], \end{aligned} \quad (4.68)$$

$$\frac{d[S_b I_{4B}]}{dt} = k_{21}[S_b I_{2B}][I_{2B}] - k_{-21}[S_b I_{4B}] + k_{25}[S_b I_{3B}][I_B] - k_{-25}[S_b I_{4B}] + k_{RB}[S_0 I_{4B}], \quad (4.69)$$

$$\frac{d[I_{2BX}]}{dt} = k_{\text{intX}}[I_B]^2 - \delta[I_{2BX}], \quad (4.70)$$

$$\frac{d[S_0 I_B]}{dt} = k_{22}[S_0][I_B] - k_{-22}[S_0 I_B] - k_{23}[S_0 I_B][I_B] + k_{-23}[S_0 I_{2B}], \quad (4.71)$$

$$\frac{d[S_0 I_{3B}]}{dt} = k_{24}[S_0 I_{2B}][I_B] - k_{-24}[S_0 I_{3B}] - k_{25}[S_0 I_{3B}][I_B] + k_{-25}[S_0 I_{4B}], \quad (4.72)$$

$$\frac{d[S_b I_B]}{dt} = k_{22}[S_b][I_B] - k_{-22}[S_b I_B] - k_{23}[S_b I_B][I_B] + k_{-23}[S_b I_{2B}], \quad (4.73)$$

$$\frac{d[S_b I_{3B}]}{dt} = k_{24}[S_b I_{2B}][I_B] - k_{-24}[S_b I_{3B}] - k_{25}[S_b I_{3B}][I_B] + k_{-25}[S_b I_{4B}], \quad (4.74)$$

$$\begin{aligned} \frac{d[S_a I_{2B}]}{dt} = & k_{20}[S_a][I_{2B}] - k_{-20}[S_a I_{2B}] + k_{-21}[S_a I_{4B}] - k_{21}[S_a I_{2B}][I_{2B}] + \\ & k_{23}[S_a I_B][I_B] - k_{-23}[S_a I_{2B}] + k_{-24}[S_a I_{3B}] - k_{24}[S_a I_{2B}][I_B], \end{aligned} \quad (4.75)$$

$$\frac{d[S_a I_{4B}]}{dt} = k_{21}[S_a I_{2B}][I_{2B}] - k_{-21}[S_a I_{4B}] + k_{25}[S_a I_{3B}][I_B] - k_{-25}[S_a I_{4B}] - k_{RB}[S_a I_{4B}], \quad (4.76)$$

$$\frac{d[S_{ab}]}{dt} = k_{-20}[S_{ab} I_{2B}] - k_{20}[S_{ab}][I_{2B}] + k_{-22}[S_{ab} I_B] - k_{22}[S_{ab}][I_B], \quad (4.77)$$

$$\begin{aligned} \frac{d[S_{ab}I_{2B}]}{dt} = & k_{20}[S_{ab}][I_{2B}] - k_{-20}[S_{ab}I_{2B}] + k_{-21}[S_{ab}I_{4B}] - k_{21}[S_{ab}I_{2B}][I_{2B}] + \\ & k_{23}[S_{ab}I_B][I_B] - k_{-23}[S_{ab}I_{2B}] + k_{-24}[S_{ab}I_{3B}] - k_{24}[S_{ab}I_{2B}][I_B], \end{aligned} \quad (4.78)$$

$$\frac{d[S_{ab}I_{4B}]}{dt} = k_{21}[S_{ab}I_{2B}][I_{2B}] - k_{-21}[S_{ab}I_{4B}] + k_{25}[S_{ab}I_{3B}][I_B] - k_{-25}[S_{ab}I_{4B}] + k_{RB}[S_aI_{4B}], \quad (4.79)$$

$$\frac{d[S_aI_B]}{dt} = k_{22}[S_a][I_B] - k_{-22}[S_aI_B] - k_{23}[S_aI_B][I_B] + k_{-23}[S_aI_{2B}], \quad (4.80)$$

$$\frac{d[S_aI_{3B}]}{dt} = k_{24}[S_aI_{2B}][I_B] - k_{-24}[S_aI_{3B}] - k_{25}[S_aI_{3B}][I_B] + k_{-25}[S_aI_{4B}], \quad (4.81)$$

$$\frac{d[S_{ab}I_B]}{dt} = k_{22}[S_{ab}][I_B] - k_{-22}[S_{ab}I_B] - k_{23}[S_{ab}I_B][I_B] + k_{-23}[S_{ab}I_{2B}], \quad (4.82)$$

$$\frac{d[S_{ab}I_{3B}]}{dt} = k_{24}[S_{ab}I_{2B}][I_B] - k_{-24}[S_{ab}I_{3B}] - k_{25}[S_{ab}I_{3B}][I_B] + k_{-25}[S_{ab}I_{4B}]. \quad (4.83)$$

The formation of intermediate DNA:protein complexes, due to monomeric and dimeric integrase binding, in our mechanistic model gives rise to multiple state variables associated with the two DNA states of interest, state 2 ( $S_a$ ) and state 4 ( $S_{ab}$ ). Summing all the ODEs describing the dynamics of state variables associated with the same DNA state of interest provides the total register of the system in those states ( $S_{aT}$  and  $S_{abT}$ ). Hence, model outputs are determined through the numerical solutions to the following ODEs:

$$\frac{dS_{aT}}{dt} = k_{RA}S_0I_{4A} - k_{RB}S_aI_{4B}, \quad (4.84)$$

$$\frac{dS_{abT}}{dt} = k_{RB}S_aI_{4B}, \quad (4.85)$$

where  $k_{RA}$  and  $k_{RB}$  are the parameters describing inversion and/or deletion mediated by integrase A and B respectively,  $S_0I_{4A}$  represents four integrase A monomers bound to DNA in state 1 and  $S_aI_{4B}$  represents four integrase B monomers bound to DNA in state 2. The model also consists of 32 parameters that are optimised through comparisons of the solutions to (4.84) and (4.85) with our experimental data.

## 4.4 Model validation via global optimisation

Induction of integrase B prior to integrase A causes transition to the unwanted deleted DNA state, state 3, and therefore we optimise our model against experimental data regarding induction of integrase A prior to integrase B only (See Experimental methods for our experimental procedure). Our data was generated as part of the research presented in [Hsiao et al., 2016], but was not presented in that publication. It is comprised of both RFP and GFP levels (state 2 and state 4 respectively) under eight distinct experimental conditions. Firstly, fluorescence is recorded for no induction of either integrase and, secondly, fluorescence is recorded for induc-

tion of integrase A only. A further six experimental procedures record fluorescence for induction of integrase B at increasing time intervals,  $\delta T$ , such that  $\delta T = 0, \dots, 5$  hours following induction of integrase A. We assume that RFP and GFP provide a direct readout of the DNA state of the system that equates to the concentration levels required to parameterise our model. Since the observed fluorescence has no physical dimension, the raw data for state 2 and state 4 are converted to percentages of the maximum fluorescence expression level recorded across all experiments to enable mathematical comparisons. This establishes the percentage fluorescence output data required to infer the parameters in our model.

Given that we are using a deterministic model to simulate recombination efficiencies within a single cell, we overcome uncertainty regarding physical quantities of DNA by allocating an initial DNA concentration (state 1) of 1, hence all model outputs are bounded within a solution space of  $[0, 1]$ . Once a numerical output has been computed, it is divided by the maximum numerical output across all simulations and multiplied by 100 in order to establish the same percentage of maximum expression captured by our converted data. Hence, model outputs are subject to the same conversion applied to the experimental data, establishing percentage changes in observed fluorescence output for varying time intervals between the induction of integrases A and B.

We employ a parallelised GA function in MATLAB on a high-performance computing cluster to perform global optimisation of the mechanistic model against our large experimental dataset. This enables us to run the GA with a large population size and over a larger number of generations within manageable time frames, and hence increases the likelihood of identifying the global optimum solution. We select a parameter space of  $[10^{-6}, 10^3]$  for all model parameters subject to inference. This is sufficiently large in light of the lack of documented reactions rate constants available in the literature, whilst minimising the potential for excessively stiff model simulations that may cause the GA to fail. We run the GA over 1000 generations to maximise the likelihood of convergence and hence identification of the global optimum solution within the parameter space. The error function is comprised of six components, each corresponding to the six datasets that we optimise the model against. Since each dataset captures percentage concentrations of varying magnitudes, an error function that computes mean absolute error consequently takes individual contributions of varying magnitudes which may skew the optimisation across all six datasets. Our error function instead computes normalised absolute error with respect to the range of each dataset:

$$\begin{aligned}
E = & \frac{1}{r_2^{NI}} \sum_{i=1}^{21} |x_{2i}^{NI} - d_{2i}^{NI}| + \frac{1}{r_4^{NI}} \sum_{i=1}^{21} |x_{4i}^{NI} - d_{4i}^{NI}| + \dots \\
& \frac{1}{r_2^A} \sum_{i=1}^{21} |x_{2i}^A - d_{2i}^A| + \frac{1}{r_4^A} \sum_{i=1}^{21} |x_{4i}^A - d_{4i}^A| + \dots \\
& \frac{1}{r_2^{AB}} \sum_{i=1}^{21} |x_{2i}^{AB} - d_{2i}^{AB}| + \frac{1}{r_4^{AB}} \sum_{i=1}^{21} |x_{4i}^{AB} - d_{4i}^{AB}|, \tag{4.86}
\end{aligned}$$

where  $E$  is the error and  $x_i$ ,  $d_i$  are the model outputs and data values at each of the twenty-one corresponding time points,  $t_i$ , respectively. The superscripts  $NI$ ,  $A$  and  $AB$  denote simulations and data corresponding to No inducer, induction of A only and induction of A and B simultaneously respectively. The subscripts 2 and 4 denote simulations and data corresponding to the DNA states of interest, state 2 ( $S_a$ ) and state 4 ( $S_{ab}$ ) respectively. The range of data values,  $r$ , is calculated for each dataset such that  $r = d_{21} - d_1$ , and is used to normalise each component of the error function.

The optimised mechanistic model is able to capture the observed system dynamics for both state 2 (Fig. 4.4A) and state 4 (Fig. 4.4B). The model is initially simulated with the parameter describing basal expression of integrases A and B ( $\gamma_{bas}$ ) simultaneously in order to generate non-zero no-inducer model outputs. Following basal simulations, the model is simulated with the parameter describing induced expression of integrases A and B ( $\gamma_{ind}$ ), but for integrase A only. Finally, the model is simulated with this induced expression parameter for integrases A and B simultaneously. The expression of fluorescent protein is greatest in state 2 for the induction of integrase A only. This is expected since the system is able to transition from state 1 to state 2 without integrase B induction causing significant competing transition to state 3. The system is able to maintain the transition to this state over time since further transitioning to state 4 is also minimal in the absence of integrase B induction; transitioning to state 3 and state 4 is only possible in this case due to basal expression of integrase B. The optimised model simulation corresponding to this case presents the greatest error observed across all six datasets, specifically with respect to the initial evolution of the response. The sigmoidal expression profile captured by this dataset is very difficult to emulate given our description of protein expression via constant parameters, and may instead require time-dependent protein expression to improve this fit. In contrast, the simultaneous induction of both integrases results in decreased fluorescence output in state 2.

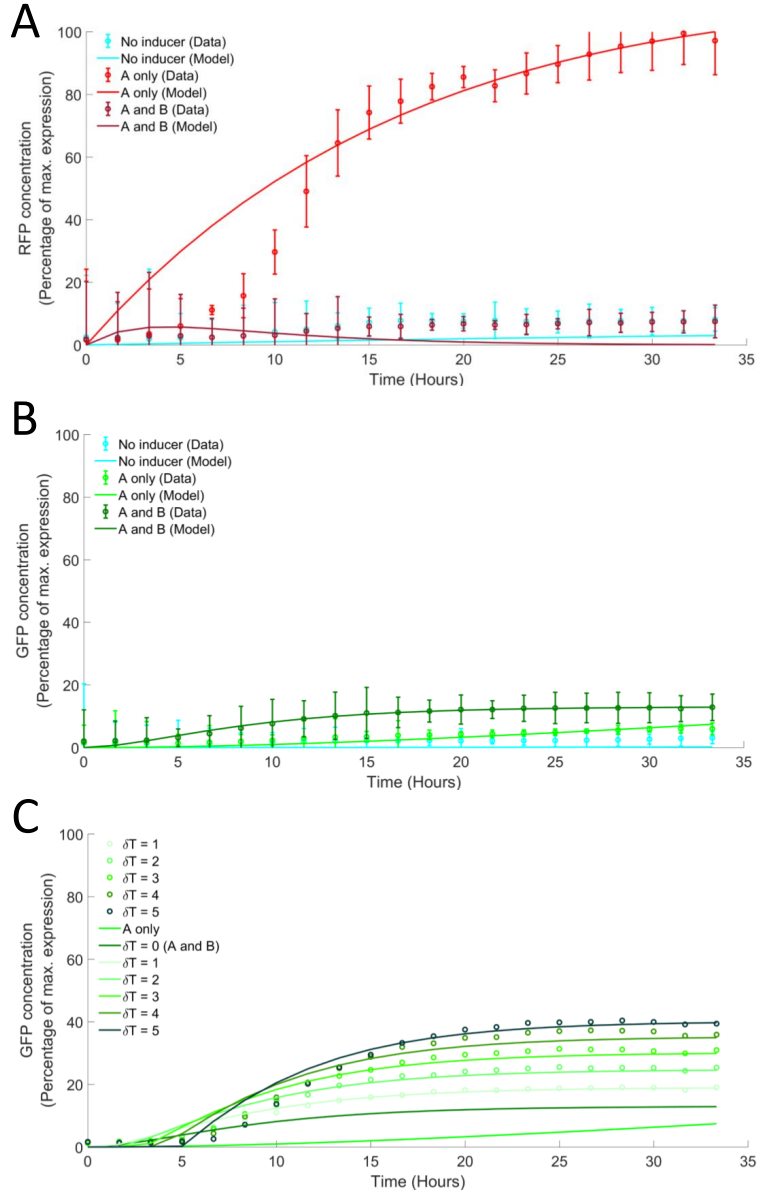


Figure 4.4: Data fitting and model prediction results. A) Optimised responses of our mechanistic model against three state 2 time course datasets. B) Optimised responses of our mechanistic model against three state 4 time course datasets. C) Model predictions of the transitioning to state 4 in response to different induction separation intervals (A only, A and B fits included from B). Circles depict experimental data; solid lines depict optimal model outputs (A only and  $\delta T = 0$ ) and model predictions ( $\delta T = 1, \dots, 5$ ). Data generated as part of the research presented in [Hsiao et al., 2016].

There is a small increase initially due to the state 2:state 3 split however, state 2 cannot be maintained since integrase B induction is able to transition this transient state to state 4. Fluorescence is relatively low in state 4 for the induction of integrase A only. This is expected since transition from state 1 to state 4, in the absence of integrase B induction, is only possible due to the basal expression of integrase B. Fluorescence increases for the simultaneous induction of both integrases, but it does not reach a similarly high level to that of state 2 for the induction of integrase A only. This is due to the fact that the transition to state 3 and state 4 occurs simultaneously and hence neither state can be maintained at maximal levels. Increased fluorescence in state 4 is thus dependent on the delay between the induction of the two integrases.

The optimal parameter set identified by the GA implies that integrase A operates on a slower time scale to that of integrase B. Although the parameters describing integrase A-mediated DNA binding interactions ( $k_{\pm 4, \dots, \pm 16}$ ) and those describing integrase B-mediated DNA binding interactions ( $k_{\pm 20, \dots, \pm 25}$ ) all take optimal values in the interval  $[10^{-2}, 10^1]$ , the parameter describing the rate of recombination mediated by integrase B,  $k_{RB}$ , is  $\sim 77\%$  greater than that of integrase A,  $k_{RA}$  (Table 4.5). These are the two key parameters in generating the numerical solutions

Parameter	Value ( $\text{M}^{-1}\text{s}^{-1}$ )	Parameter	Value ( $\text{s}^{-1}$ )	Parameter	Value ( $\text{Ms}^{-1}$ )
$k_1$	0.1178	$k_{-1}$	0.2507	$\gamma_{\text{bas}}$	0.0311
$k_4$	0.6046	$k_{-4}$	0.9999	$\gamma_{\text{ind}}$	0.0988
$k_5$	1.1962	$k_{-5}$	1.2614	—	—
$k_{13}$	1.5835	$k_{-13}$	0.6076	—	—
$k_{14}$	0.2511	$k_{-14}$	0.0201	—	—
$k_{15}$	0.1693	$k_{-15}$	1.0903	—	—
$k_{16}$	0.2192	$k_{-16}$	0.9371	—	—
$k_{\text{intX}}$	1.4968	$k_{RA}$	0.4239	—	—
$k_{20}$	0.7169	$k_{-20}$	0.2434	—	—
$k_{21}$	0.1541	$k_{-21}$	0.8693	—	—
$k_{22}$	2.6537	$k_{-22}$	0.3342	—	—
$k_{23}$	0.9586	$k_{-23}$	0.5976	—	—
$k_{24}$	0.3745	$k_{-24}$	1.0478	—	—
$k_{25}$	1.0276	$k_{-25}$	4.1887	—	—
—	—	$k_{RB}$	0.7535	—	—
—	—	$\delta$	0.6401	—	—

Table 4.5: The optimal model parameter values inferred by the genetic algorithm through global optimisation. Model parameters are dimensional, taking SI units arising from standard mass action kinetics.

to (4.84) and (4.85) that comprise our model simulations and hence it appears that the rate of inversion/deletion mediated by integrase B is greater than that of integrase A in producing the dynamical behaviour captured by our experimental data. This implies that the action of integrase B is naturally faster than that of integrase A and may therefore be more useful as a component in the design of synthetic bi-



ological circuitry. Note that we have assumed the parameters describing protein binding interactions ( $k_{\pm 1}$ ,  $k_{\text{intX}}$ ) are identical for each of the two integrases, just as the protein expression and degradation parameters. This minimises the overall number of model parameters, increasing the transferability of the model, and facilitates the inference of functional distinctions between the two integrase inputs as described above. Further examination of potential protein binding distinctions is possible for future studies at the cost of increasing the number of model parameters. The optimal parameter set is used to generate all plots in Fig. 4.4, as well as all subsequent plots.

After training our model using the datasets in Fig. 4.4A and 4.4B, we used our optimised mechanistic model to predict a set of experimental data that was excluded from the training set. Optimised model predictions align with experimental data for increasing integrase induction delays ( $\delta T = 1, \dots, 5$ ) (Fig. 4.4C). As the delay between the induction of integrase A and B increases, the time afforded to the maintenance of state 2 also increases. There is therefore a greater concentration of state 2 than state 1 when integrase B is eventually induced and hence transition to state 4 is increased by virtue of integrase B-mediated inversion. Consequently, the transition to state 3 is decreased due to the decrease in concentration of state 1.

Additionally, we validated our model by predicting endpoint GFP concentrations relating to both A then B temporal response data and an entirely separate dataset regarding B then A temporal responses. The endpoint response of the system as a function of the integrase induction separation interval  $\delta T$  is shown in Fig. 4.5. Optimal endpoint percentage model outputs as a function of  $\delta T$  align closely with that of the experimental data, providing further evidence of the model’s predictive capability. Fig. 4.5 also supports the notion that the efficiency of recombination induction via integrase B must be superior to that of integrase A since identical efficiencies would be expected to result in a 50:50 split for  $\delta T = 0$ . This inequality in integrase-mediated inversion was previously observed in [Hsiao et al., 2016], but no mechanistic comparisons had been performed at that time. Consequently, it remains to experimentally examine functional differences between distinct integrases as these properties may allow for specific logic operations dependent on the pair of integrases selected, the arrangement of the associated attachment sites and the specific roles each integrase input is assigned in the circuit.

#### 4.4.1 Experimental methods

The experimental system for the temporal logic gate with Bxb1 and TP901-1 integrases was implemented in DH5a-Z1 *E. coli*. The Bxb1 and TP901-1 integrases are

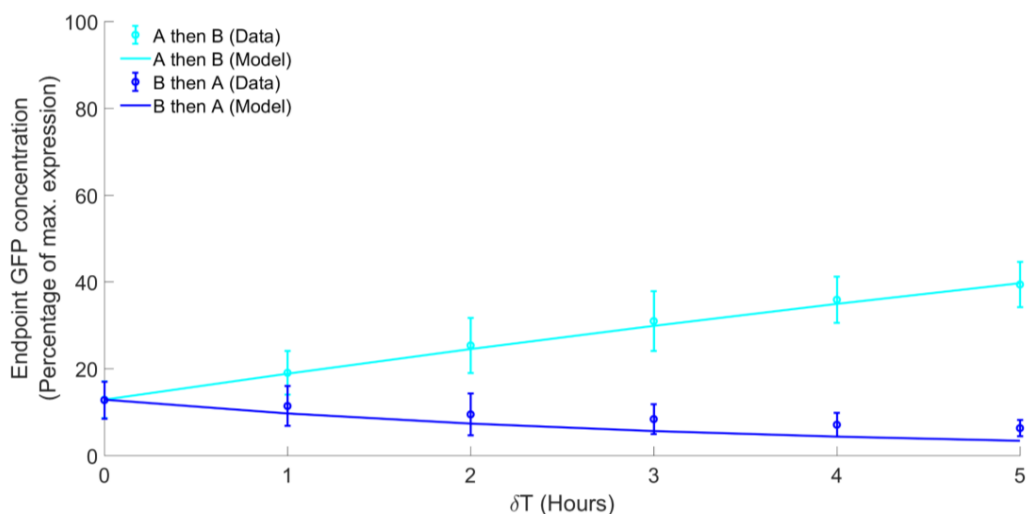


Figure 4.5: Endpoint GFP concentration results. A then B GFP percentage concentration endpoint predictions from Fig. 4.4C plotted as a function of  $\delta T$  ( $\delta T = 0, \dots, 5$ ; light blue plot line). Equivalent predictions of B then A temporal responses are depicted by the dark blue plot line. Model simulations generated using our optimal parameter set. Data generated as part of the research presented in [Hsiao et al., 2016].

on a high copy plasmid (available from the Addgene plasmid repository, ID 82351). The temporal logic gate with integrase binding targets was chromosomally integrated into the Phi80 site of the *E. coli* genome using CRIM integration [Haldimann and Wanner, 2001]. A plasmid version of the same logic gate is also available from Addgene (ID 82352).

M9CA media was prepared with  $1 \times$  M9 salts (Teknova, M1906) augmented with 100 mM  $\text{NH}_4\text{Cl}$ , 2 mM  $\text{MgSO}_4$ , 0.01% casamino acids,  $0.15 \mu\text{g/ml}$  biotin, and  $1.5 \mu\text{M}$  thiamine. 0.2% glycerol was used as the sole carbon source and the entire solution was sterile-filtered ( $0.2 \mu\text{m}$ ). During the experiment, all media contained the antibiotics chloramphenicol (Sigma-Aldrich, Inc (C0378);  $50 \mu\text{g/ml}$ ) and kanamycin (Sigma-Aldrich, Inc (K1876);  $30 \mu\text{g/ml}$ ). L-arabinose, the inducer for TP901-1, was used at a concentration of 0.01% by volume, and anhydrous tetracycline (aTc), the inducer for Bxb1, was used a concentration of  $200 \text{ ng/ml}$  ( $450 \text{ nM}$ ). All experiments were performed with the aid of timed liquid handling by a Hamilton STARlet Liquid Handling Robot (Hamilton Company).

At the beginning of the experiment, the cells were diluted to OD 0.06-0.1 into a 96-well matriplate (Brooks Automation, Inc., MGB096-1-2-LG-L) with  $500 \mu\text{l}$  total volume in M9CA. Cultures were incubated at  $37^\circ\text{C}$  in a BioTek Synergy H1F

plate reader with linear shaking (1096 cycles per minute) (BioTek Instruments, Inc.), and inducers were added at various  $\delta T$  time points by the Hamilton robot. OD and fluorescence measurements (superfolder-GFP ex488/em520, mKate2-RFP ex580/em610) were taken every 10 min. Each experimental condition was performed on the plate in triplicate.

## 4.5 Reversing the roles of integrase inputs

By allowing integrase B to occupy the role of integrase A and *vice versa*, we are able to investigate the effect of input reversal on the output of the temporal logic gate *in silico* using our optimised mechanistic model. This involves swapping the optimal parameter values corresponding to reactions mediated by integrase A with the equivalent optimal parameter values corresponding to reactions mediated by integrase B. That is, our previous ‘A then B’ inputs are reversed in order to examine ‘B then A’ simulations. We simulate the same dynamical responses captured by our experimental data (Fig. 4.6).

The effect of reversing integrase inputs on state 2, for the induction of integrase B only, results in an increased response time that results in faster transitioning to maximal RFP concentration (Fig. 4.6A). Induction of both integrases ( $\delta T = 0$ ) results in a dynamical response that exhibits a significant increase in transient RFP concentration compared to the original inputs. The expected sequestration of RFP concentration can be observed, however it is unable to reach 0 on the same timescale as the original input data. These simulations demonstrate the increased speed of the reactions mediated by integrase B since this transient state is expressed to a greater level before sequestration via the slower action of integrase A. Input reversal also has a significant impact on the expression of state 4 (Fig. 4.6B). Although the induction of integrase B only causes increased transition to state 2, the relatively slower action of integrase A results in negligible GFP concentration. However, simultaneous induction of both integrases ( $\delta T = 0$ ) causes a significant increase in concentration on the same timescale as the original inputs, thus demonstrating that the action of the slowest integrase in the input pair is rate limiting in the overall dynamical response of the temporal logic gate. This increase in concentration is likely due to the increased concentration of state 2 caused by integrase B which can be transitioned to state 4 via integrase A.

Induction separation intervals provide further evidence of the increased concentration achieved by reversing the integrase inputs (Fig. 4.6C). Here, the endpoint responses as a function of  $\delta T$  confirm that the expected 50% endpoint concentration

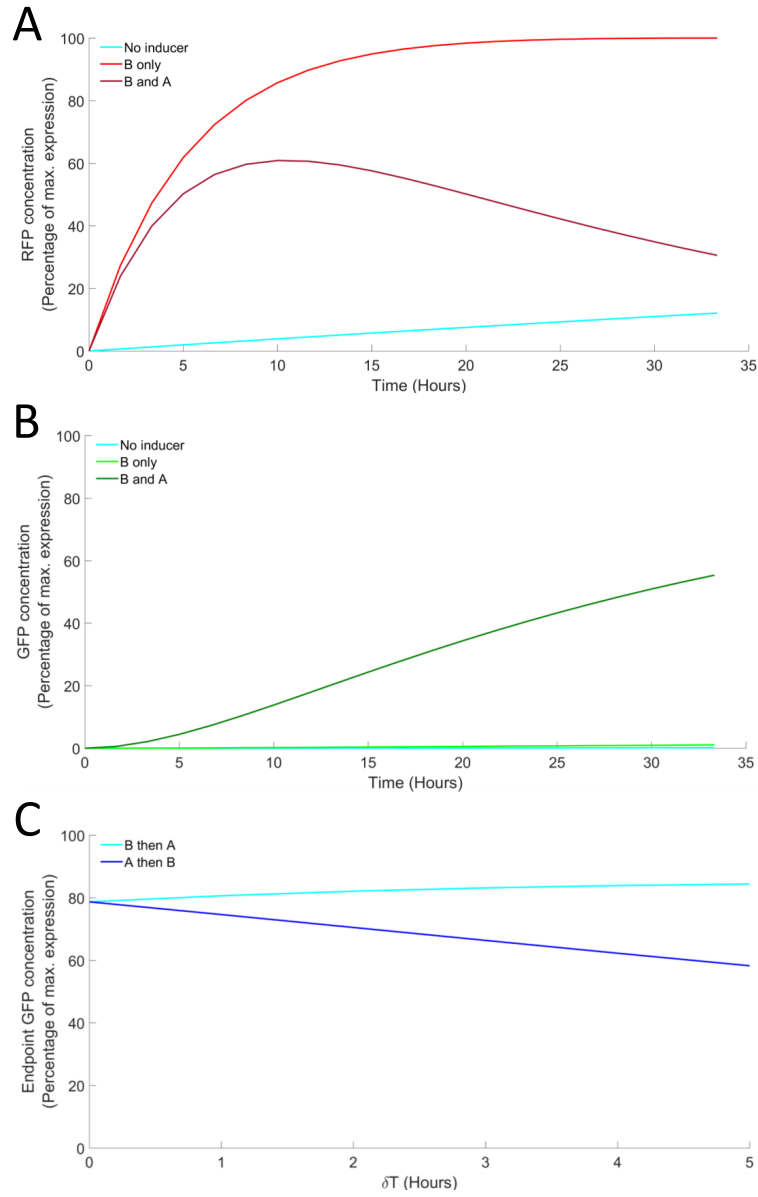


Figure 4.6: Model simulations of reversed integrase inputs. A) Optimised model simulations of state 2 RFP concentration using reversed integrase inputs (No inducer, B only, B and A ( $\delta T = 0$ )). B) Optimised model simulations of state 4 GFP concentration using reversed integrase inputs (No inducer, B only, B and A ( $\delta T = 0$ )). C) B then A GFP percentage concentration endpoint simulations plotted as a function of  $\delta T$  ( $\delta T = 0, \dots, 5$ ; light blue plot line). Equivalent predictions of A then B temporal responses are shown by the dark blue plot line.

for  $\delta T = 0$  is shifted significantly towards the concentration of state 4 and the expression of GFP ( $\sim 80\%$ ). This highlights the potential benefit of employing the more efficient integrase in the original role of integrase A, given the decreased transitioning to the unwanted state 3 that this provides. The endpoint concentrations plotted in Fig. 4.6C are taken from outputs simulated over an increased timescale in order to obtain steady-state levels that are not reached on the original timescale. As such, the increased efficiency of the circuit to transition to state 4 requires more time, and hence it is likely that a suitable trade-off between response time and efficiency will be required for optimal performance. The speeds at which the system is able to transition to each distinct DNA state for both input assignments are summarised in Fig. 4.7.

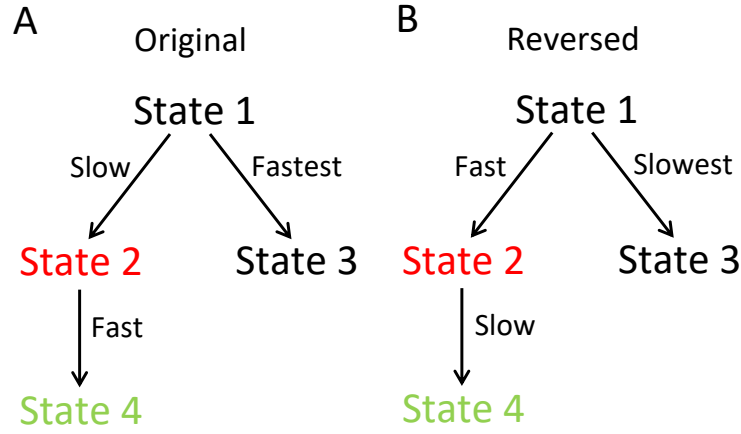


Figure 4.7: State transition speeds associated with input assignment. A) The original circuit exhibits a faster transition to state 4 following slower accumulation of state 2. B) The reversed circuit exhibits a slower transition to state 4 following faster accumulation of state 2.

## 4.6 Conclusions

We have developed the first mechanistic mathematical model of a synthetic two-input temporal logic gate using previously validated models of *in vitro* DNA recombination reactions. Our model was validated against a series of time course datasets, demonstrating quantitative replication of *in vivo* datasets relating to the induction of none, one, or both integrases, and accurate prediction of the response of the logic gate to inputs separated by five different induction separation intervals. Further error reduction relating to the rise time of fluorescence output may be possible by accounting for increased detail regarding the nature of integrase expression, however

the current model provides sufficient replication of the corresponding steady state outputs which has greater importance as feature of a reliable design tool. Both our modelling investigation and our experimental data provide evidence of functional distinctions between the two integrase inputs, suggesting that integrase B, Bxb1, operates more efficiently than integrase A, TP901-1 ( $\sim 1.8$ -fold faster). Experimental data for other distinct serine integrases would allow us to develop a lookup table of logic functions dependent on the choice and assignment of inputs. The effect of reversing the roles of the two integrases was subsequently shown to elicit a more rapid response time as well as significantly greater GFP expression in state 4. Mechanistic models therefore have the potential to reveal functional nuances that might exist between other characterised integrases and hence inform further experimental verification. These results could therefore also have important implications in the design of higher-order logic circuitry when considering the ideal pairs of integrase inputs to select in order to realise the desired system output. Future work will extend our modelling investigations to higher-order logic circuits, namely 2-4 decoders in mammalian cells for potential therapeutic applications.

## Chapter 5

# DNA Recombination Experiments

### 5.1 Establishing ideal bacterial growth conditions

In order to develop our biological understanding of recombinase-based genetic switches, we performed experiments designed to facilitate the collection of primary data on the efficiency of bacterial DNA recombination *in vivo*. We implemented an integrase-excisionase-mediated genetic switch in the bacteria *Streptomyces coelicolor* by integrating an operon consisting of five *lux* genes, *luxCDABE*, into the *S. coelicolor* M145 genome in order to enable the cells to exhibit luminescence. The *luxA* and *luxB* genes produce luciferase enzymes that catalyse bioluminescence reactions, and the three remaining genes produce the enzymes that provide the substrate of these reactions. The production of the RDF gp3 in conjunction with the natural production of pSAM2 integrase should be sufficient to cause excision of the *luxCDABE* gene from the bacterial genome and should therefore ‘turn off’ luminescence. In theory, the efficiency of the pSAM2-gp3-mediated excision of the *luxCDABE* operon can be quantified by analysing the luminescence output of the bacteria.

The *luxCDABE* operon is carried on the L3 vector (Figure 5.1A). All *S. coelicolor* M145 strains cultured in our experiments had the L3 vector integrated into their genome. Integration of the L3 vector also resulted in resistance to the antibiotic apramycin which allowed isolation of clones with the insert of interest. In addition, these strains also had either the pCC4 vector (Figure 5.1B) or pRDF1 vector (Figure 5.1C) integrated into their genome. The pCC4 vector is ‘empty’ in the sense that it does not contain the gene *gp3* necessary for the production of gp3 excisionase, but it does provide resistance to the antibiotics apramycin and

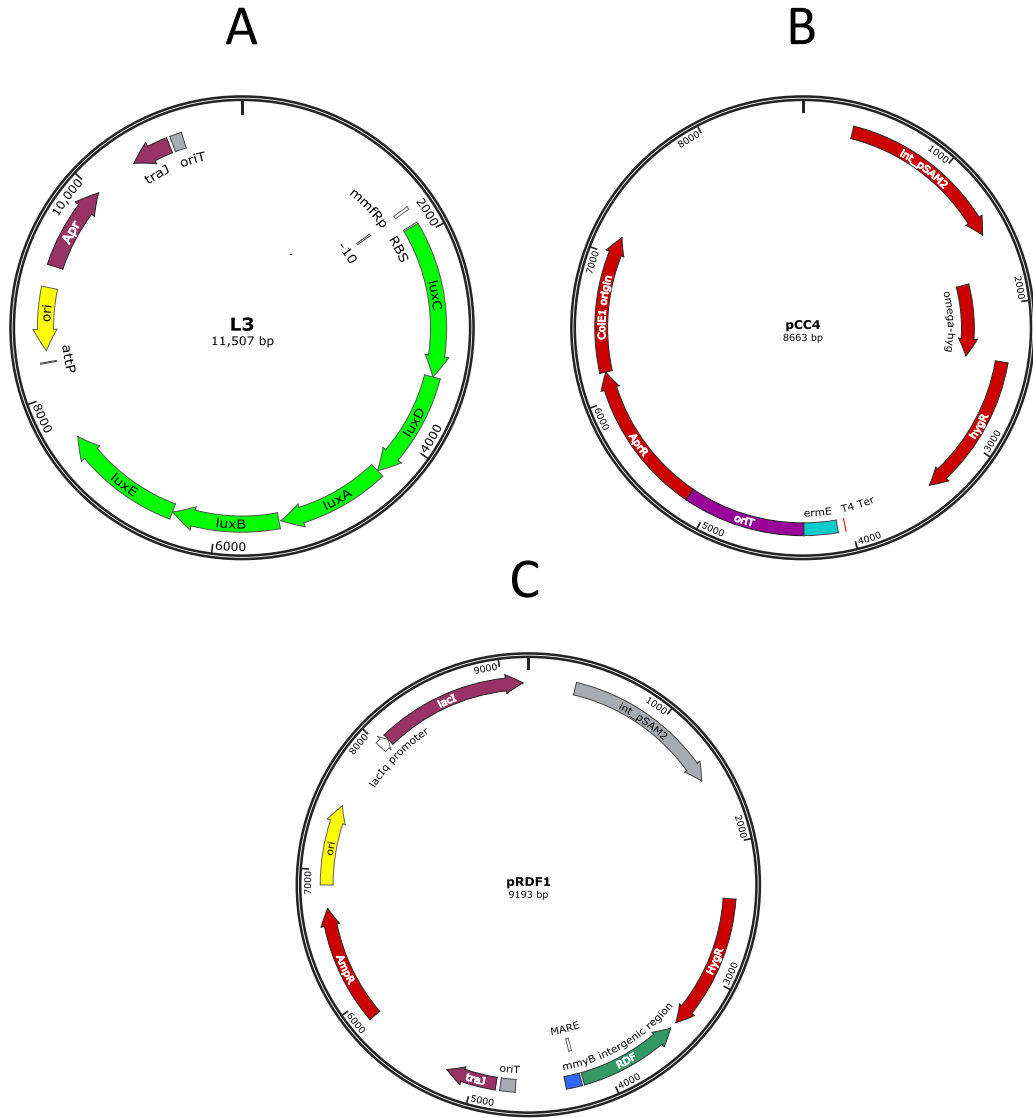


Figure 5.1: Schematics of the vectors used in our DNA recombination experiments. A) The L3 vector containing the *luxCDABE* operon and the apramycin resistance gene. B) The pCC4 vector containing apramycin and hygromycin resistance genes and the pSAM2 integrase gene. C) The pRDF1 vector containing ampicillin and hygromycin resistance genes, the pSAM2 integrase gene and the RDF gene, *gp3*.

hygromycin and also naturally produces pSAM2 integrase (Table 5.1A). We refer to *S. coelicolor* M145 strains containing the L3 and pCC4 vectors as L3+pCC4 and these strains therefore provided an excisionase-less control (Table 5.1B). The pRDF1 vector is identical to pCC4 with the exception that it does contain the *gp3* gene and provides resistance to ampicillin instead of apramycin. We refer to *S. coelicolor*



A		
Vector	Genes of interest	Resistance
L3	<i>luxCDABE</i>	Apr
pCC4	-	Apr and Hyg
pRDF1	<i>gp3</i>	Amp and Hyg

B		
Strain	Genes of interest	Resistance
L3+pCC4	<i>luxCDABE</i>	Apr and Hyg
L3+pRDF1	<i>luxCDABE</i> and <i>gp3</i>	Apr, Amp and Hyg

Table 5.1: A) Summary of the vectors used in our DNA recombination experiments. B) The nomenclature relating to *S. coelicolor* M145 strains containing these vectors. Apr, Hyg and Amp denote the antibiotics apramycin, hygromycin and ampicillin respectively.

M145 strains containing the L3 and pRDF1 vectors as L3+pRDF1.

The integration of the *luxCDABE* operon into the *S. coelicolor* M145 genome is the ‘on’ switch in this system. To carry out DNA transfer (integration), *E. coli* ET12567 cells containing the *luxCDABE* operon within the plasmid L3 vector are grown to an optimal growth phase in liquid culture, then centrifuged and re-suspended to remove any antibiotics used in the culture. *S. coelicolor* cells are heat shocked for 10 minutes to make them more likely to accept the new DNA. Both the ET12567 cells and *S. coelicolor* are plated out on soya flour mannitol (SFM) plates and left to grow overnight (see Methods for SFM protocol). It is during this time that the transfer of genetic material should occur. This is known as intergenic conjugation (see Methods for full protocol). The following day the plates are overlaid with nalidixic acid, which should kill all *E. coli*, and apramycin which should kill any *S. coelicolor* which do not contain the desired genetic insert, thus selecting for the *luxCDABE* operon. This process was also carried out for each of the pCC4 and pRDF1 vectors, with hygromycin used to select for the desired strains rather than apramycin. This does restrict our ability to test the efficiency of the integration reaction since only the *S. coelicolor* strains with the desired integrated genes are preserved. However, since the integration reaction is mediated purely by integrase, integration is thought to be highly efficient in the absence of RDF, a notion that is supported by our mathematical modelling investigation. We are therefore interested in how efficiently the RDF is able to excise *luxCDABE* out of the *S. coelicolor* genome in conjunction with pSAM2 integrase, turning the switch ‘off’ again and thus rendering the bacteria incapable of producing luminescence.

Before examining the efficiency of the excision reaction, we performed a number of experiments designed to elucidate the ideal growth conditions for *S. coelicolor* that permit optimal luminescence production in the integrated cells. These preliminary trials were carried out using L3+pCC4 (containing no gp3) that is, we wanted to optimise the protocol without RDF activity presenting additional variables. We initially plated out L3+pCC4 strains onto 10 SFM plates, directly from glycerol stocks. Our SFM plates were produced by adding 25 mL of SFM to a standard Petri dish using a sterile pipette. The bacteria were diluted according to a serial dilution that gave rise to a countable number of colonies; highly concentrated bacterial solutions grow very prolifically on SFM plates, forming ‘lawns’ of cells that make it impossible to distinguish individual colonies. Each specific dilution is spread onto its own SFM plate to permit the measurement of luminescence for each individual colony. Serial dilution involves pipetting an appropriate volume of stock solution (100%) into a volume of sterile water sufficient to produce a concentration of 10% solution. For example, 20  $\mu$ L of solution into 180  $\mu$ L of sterile water gives 20  $\mu$ L of solution in a total 200  $\mu$ L of mixture and is therefore 10% diluted. By pipetting 20  $\mu$ L of the 10% dilution into another 180  $\mu$ L of sterile water, the dilution is increased further by a factor of 10 giving a 1% diluted mixture. Repeating this process produces a series of diluted solutions each of which is one tenth of the concentration of the preceding solution and is an effective method of establishing the appropriate concentrations that give rise to countable numbers of bacterial colonies. To measure luminescence, we utilised a charge coupled device (CCD) camera which counts the photons emitted via bioluminescence, and Image32 computer software which produces images of the analysed plate that depict the luminescence of colonies in their exact locations on the plate.

All colonies grown on the 10 initial plates were expected to exhibit luminescence however, only 77 out of 94 colonies ( $\sim 82\%$ ) were seen to do so (Table 5.2). It is possible that this low percentage was caused by the freezing of glycerol stocks and therefore, in an attempt to improve this result, we carried out a repeat experiment in conjunction with a trial of an alternative method. We picked three single colonies that had already been seen to exhibit luminescence on an SFM plate and serially diluted these in order to achieve a countable number of colonies. Picking colonies that had already exhibited luminescence was thought to increase the likelihood of maintaining luminescence. The resulting dilution was spread on 3 SFM plates. The repeat produced a result consistent with the unexpected result of the first trial, with 84 of 99 colonies ( $\sim 85\%$ ) exhibiting luminescence (Table 5.3A). The alternative method, whereby luminescent colonies were picked from a ‘starter

Plate number	Total colonies	Luminescent colonies
1	12	12
2	12	11
3	16	11
4	11	5
5	10	8
6	9	8
7	6	5
8	8	7
9	7	7
10	3	3
Total	94	77

Table 5.2: Luminescence in L3+pCC4, grown from glycerol stocks.

A			B		
Plate number	Total colonies	Luminescent colonies	Plate number	Total colonies	Luminescent colonies
1	22	20	1	26	26
2	17	14	2	15	14
3	13	11	3	8	8
4	23	17	Total	49	48
5	24	22			
Total	99	84			

Table 5.3: A) Luminescence in L3+pCC4, grown from glycerol stocks (repeat). B) Luminescence in L3+pCC4, grown from a dilution of luminescent colonies (starter plate).

plate’, did however produce the result we expected with only one colony identified as not exhibiting luminescence ( $\sim 100\%$ ) (Table 5.3B). This indicated that the freezing process for glycerol stocks may have a detrimental effect on luminescence.

A repeat experiment using this starter plate method was carried out to validate the result. In this case, six single colonies that had exhibited luminescence were picked and spread onto six SFM plates at the appropriate dilution to produce single colonies. These colonies gave rise only to luminescent colonies and hence this growth method is most efficient in preserving the expression of the *luxCDABE* operon in *S. coelicolor* (Table 5.4). It appears that plating *S. coelicolor* strains directly from glycerol stocks is conducive to unpredictable results, most likely due to the freezing of the cells for storage purposes.

Plate number	Total colonies	Luminescent colonies
1	14	14
2	10	10
3	2	2
4	24	24
5	6	6
6	10	10
Total	66	66

Table 5.4: Luminescence in L3+pCC4, grown from a starter plate (repeat).

We performed an additional experiment to examine the luminescence of lawns of L3+pCC4. One luminescent colony and four non-luminescent colonies were picked and spread onto standard Petri dishes of 25 mL SFM. The lawn grown from the luminescent bacteria gave at least a 50-fold increase in the light reading compared to that produced by the four lawns of previously non-luminescent bacteria (Table 5.5). These four lawns were not expected to exhibit luminescence however, even allowing

Plate number	Colony of origin	Light reading
1	Luminescent	3109779
2	Not luminescent	955
3	Not luminescent	58485
4	Not luminescent	17688
5	Not luminescent	609

Table 5.5: Light readings taken from five lawns of L3+pCC4 using a CCD camera and Image32 software.

for background noise, plates 3 and 4 in particular appear to exhibit unexpectedly high light readings. It appears that these cells did, in fact, retain the *lux* genes which makes it unclear why they did not exhibit luminescence initially. It might be that these particular bacteria require a period of time to adopt regular cellular functionality that has skewed our preliminary results.

Having established that L3+pCC4 cultures originating from luminescent colonies provide the most suitable test cultures, we were able to run trials relating to the activity of gp3 excisionase in this genetic switch system.

## 5.2 Recording excision time courses

In order to investigate the efficiency of the excision reaction, RDF must be produced within the system. We therefore performed a series of experiments implementing L3+pRDF1 strains that possess both the *luxCDABE* operon and *gp3* gene in order to record time course excision efficiency. We aimed to record luminescence at four time points (0, 24, 48 and 72 hours) to establish the time course evolution of excision efficiency. These time points were chosen based on optimisation trials run by an experimental collaborator working with similar strains [Styles, 2016]. We decided to grow our L3+pRDF1 strains in liquid media 2xYT to facilitate the plating of bacterial solutions at the intended time points (see Methods for 2xYT protocol). Growing bacterial cultures in liquid media causes colonies to develop in clumps to an extent that some are too large to be effectively taken up by a pipette. As a result, we grew the cultures with glass beads mixed in with the solution to prevent aggregation, the accumulation of large clumps. A total of twenty beads were added to each flask and all flasks were incubated at 30 °C in a shaking incubator at 200 rpm. We spread our liquid samples onto solid SFM plates at the aforementioned time points, however the light readings were taken 72 hours after each sample was plated to allow enough time for colonies to develop. We required 72 hours to achieve sufficient growth for measurable luminescence levels due to the doubling time of *S. coelicolor* [Chen and Qin, 2011]. Therefore, the time points at which we recorded the luminescence of the L3+pRDF1 colonies were in fact 0(+72), 24(+72), 48(+72) and 72(+72) hours.

The first time point was recorded immediately after the appropriate cultures were inoculated in order to get a 0 hour time point reading. L3+pRDF1 strains were plated on 25 mL SFM plates; one non-selective set without antibiotics and one selective set with 12.5 µg/mL hygromycin and 12.5 µg/mL apramycin. The selective plates select for *gp3* and *luxCDABE* respectively, thus the only colonies that should grow successfully on this medium are those which have retained all integrated genes, including the resistance markers. Any cells which have had their *luxCDABE* operon excised are expected to have also lost the associated apramycin resistance and should therefore not grow on these plates. Hence, we expected the RDF *gp3* to take effect on both selective and non-selective SFM plates but we expected fewer colonies, that are all luminescent, to grow on the selective plates (Fig. 5.2). This process was carried out for three different dilutions (Table 5.6) to increase the chance of producing countable numbers of colonies. All experiments were also performed using L3+pCC4 strains as an experimental control. That is, bioluminescence was expected in all the

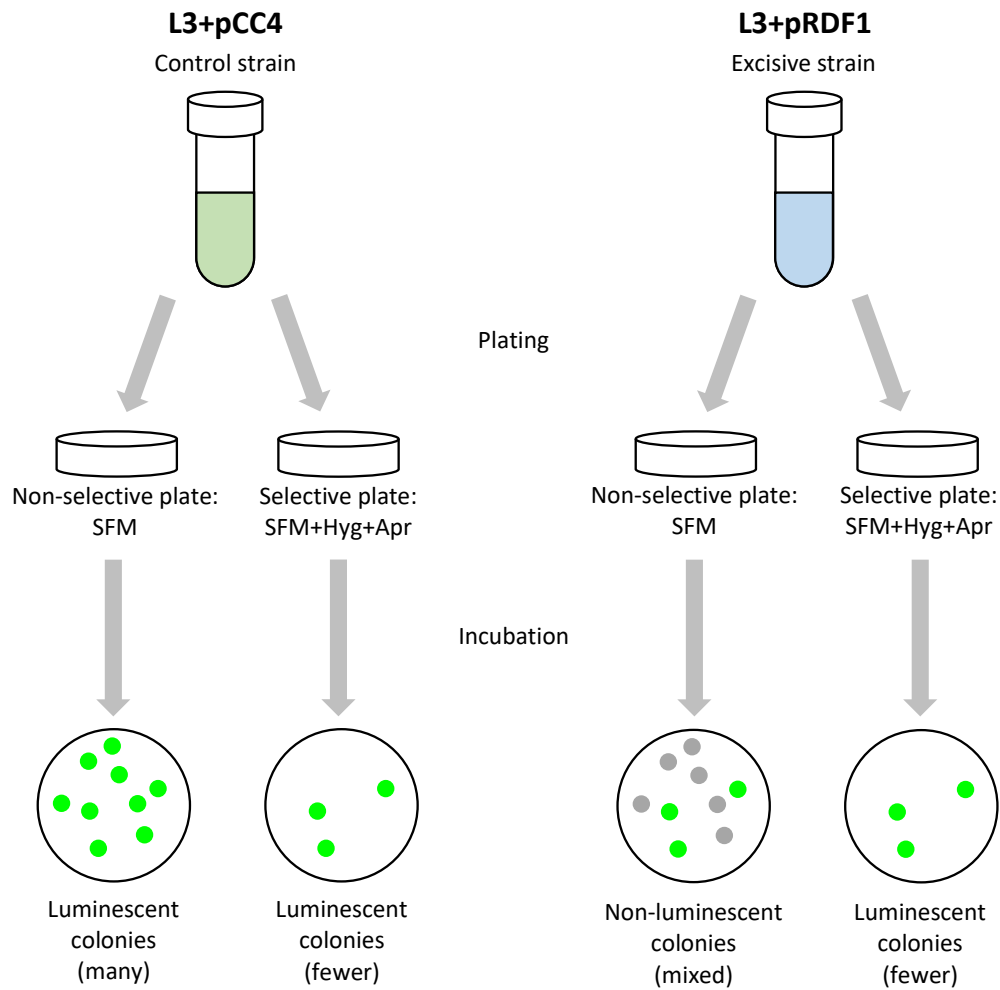


Figure 5.2: Experimental design. The control strain L3+pCC4 and the excisive strain L3+pRDF1 were used to record excision efficiency. L3+pCC4 strains on non-selective media are expected to produce many luminescent colonies since there is no gp3 to excise the *luxCDABE* operon. L3+pCC4 strains on selective media are also expected to produce luminescent colonies, but in lower numbers due to antibiotic selection which kills colonies of strains that do not have the full complement of resistance markers. L3+pRDF1 strains on non-selective media are expected to produce non-luminescent colonies since the presence of gp3 is sufficient to mediate excision of the *luxCDABE* operon in conjunction with pSAM2 integrase; the efficiency of this reaction is determined by the number of non-luminescent colonies. L3+pRDF1 strains on selective media are expected to produce low numbers of luminescent colonies since excision of the *luxCDABE* operon removes apramycin resistance and hence only strains that retain the operon are preserved; a fully efficient excision reaction would result in no growth. Hyg and Apr denote hygromycin and apramycin respectively.

0 hours (+72 hours)	L3+pRDF1			
	SFM		SFM+Hyg+Apr	
Dilution	Col	Lum	Col	Lum
10 <sup>0</sup>	Lawn	Even	Lawn	Even
10 <sup>-1</sup>	TMTC	Even	42	Even
10 <sup>-2</sup>	23	19	6	6

	L3+pCC4			
	SFM		SFM+Hyg+Apr	
Dilution	Col	Lum	Col	Lum
10 <sup>0</sup>	Lawn	Even	Lawn	Even
10 <sup>-1</sup>	TMTC	Even	TMTC	Even
10 <sup>-2</sup>	TMTC	Even	TMTC	Even

Table 5.6: Excision efficiency at 0 hours. The selective media is denoted by SFM+Hyg+Apr where Hyg and Apr denote 12.5  $\mu\text{g/mL}$  of hygromycin and apramycin respectively. The abbreviations Col and Lum denote the total number of colonies and the number of those colonies that exhibited luminescence respectively. Even denotes even luminescence observed across the plate. TMTC is an abbreviation of too many to count.

control experiments since these strains were void of any integrated *gp3* gene and were therefore incapable of pSAM2-*gp3*-mediated excision. As expected, the L3+pCC4 strains grew very efficiently on the selective media at the chosen dilutions and all of these colonies, or lawns of colonies, exhibited luminescence. The L3+pRDF1 strains grew in smaller numbers at this 0 hour time point and all of these colonies also exhibited luminescence with the exception of four on the  $10^{-2}$  non-selective SFM plate. This suggests that the *gp3*, in conjunction with pSAM2, had excised the *luxCDABE* gene from these strains and was therefore functioning as expected, with an efficiency of  $\sim 17\%$ .

After the 0 hour collection, subsequent measurements were recorded every 24 hours, with some adaptations. We decided to increase the number of dilutions to be plated in light of the growth recorded at 0 hours and also added different amounts of hygromycin and apramycin to some of the dilutions in order to optimise the selection of the desired strains. Since the L3+pCC4 strains were seen to grow more efficiently than the L3+pRDF1 strains, we diluted these solutions further to produce countable numbers of colonies for both strains. Bacterial growth occurred on half of the plates used for the 24 hour time point (Table 5.7), with the vast majority of colonies exhibiting luminescence. Again, all L3+pCC4 colonies were luminescent, as expected. The two L3+pRDF1 colonies on selective media that

24 hours (+72 hours)	Dilution	L3+pRDF1			
		SFM		SFM+Hyg+Apr	
		Col	Lum	Col	Lum
$10^{-3}$	Hyg+2Apr	23	19	8	7
$10^{-3}$	$\frac{1}{2}$ Hyg+ $\frac{1}{2}$ Apr	8	7	4	3
$10^{-4}$	$\frac{1}{2}$ Hyg+ $\frac{1}{2}$ Apr	-	-	-	-
$10^{-5}$	$\frac{1}{2}$ Hyg+ $\frac{1}{2}$ Apr	-	-	-	-
$10^{-5}$	No antibiotic	6	6	1	1
$10^{-6}$	No antibiotic	-	-	-	-
$10^{-7}$	No antibiotic	3	3	-	-

	Dilution	L3+pCC4			
		SFM		SFM+Hyg+Apr	
		Col	Lum	Col	Lum
$10^{-8}$	Hyg+2Apr	18	18	2	2
$10^{-7}$	$\frac{1}{2}$ Hyg+ $\frac{1}{2}$ Apr	6	6	4	4
$10^{-8}$	$\frac{1}{2}$ Hyg+ $\frac{1}{2}$ Apr	-	-	1	1
$10^{-9}$	$\frac{1}{2}$ Hyg+ $\frac{1}{2}$ Apr	6	6	28	28
$10^{-7}$	No antibiotic	-	-	-	-
$10^{-8}$	No antibiotic	-	-	-	-
$10^{-9}$	No antibiotic	-	-	-	-

Table 5.7: Excision efficiency at 24 hours. The selective media is denoted by SFM+Hyg+Apr where Hyg and Apr denote 12.5  $\mu\text{g/mL}$  of hygromycin and apramycin respectively. The 12.5  $\mu\text{g/mL}$  concentration is doubled or halved to 25  $\mu\text{g/mL}$  or 6.25  $\mu\text{g/mL}$  respectively as denoted in the table. The abbreviations Col and Lum denote the total number of colonies and the number of those colonies that exhibited luminescence respectively. Dashes denote plates on which no growth occurred and therefore could not give a light reading.

did not exhibit luminescence indicate that the *luxCDABE* genes were not being expressed properly since all bacteria void of this gene should have been killed. The five non-luminescent L3+pRDF1 colonies that grew on non-selective media with antibiotic added to their dilution was an unexpected result given that the equivalent colonies were all luminescent in collections from the 0 hour time point. Although there were fewer colonies on the selective plates which suggests that the gp3 was active, the nine colonies on non-selective SFM plates with no antibiotic in their dilution were all luminescent which contradicts our observations at the 0 hour time point and thus suggests that the gp3 was, in fact, not having an effect.

The number of dilutions used was increased further for the 48 hour time point. Bacterial growth occurred on 12 of the 36 plates (Table 5.8) and, again, the



48 hours (+72 hours)		L3+pRDF1			
		SFM		SFM+Hyg+Apr	
Dilution		Col	Lum	Col	Lum
$10^{-4}$	Hyg+2Apr	-	-	-	-
$10^{-3}$	$\frac{1}{2}$ Hyg+ $\frac{1}{2}$ Apr	4	3	2	2
$10^{-4}$	$\frac{1}{2}$ Hyg+ $\frac{1}{2}$ Apr	2	2	-	-
$10^{-5}$	$\frac{1}{2}$ Hyg+ $\frac{1}{2}$ Apr	-	-	-	-
$10^{-6}$	$\frac{1}{2}$ Hyg+ $\frac{1}{2}$ Apr	-	-	-	-
$10^{-5}$	No antibiotic	2	2	-	-
$10^{-6}$	No antibiotic	-	-	-	-
$10^{-7}$	No antibiotic	-	-	-	-
$10^{-8}$	No antibiotic	-	-	-	-

		L3+pCC4			
		SFM		SFM+Hyg+Apr	
Dilution		Col	Lum	Col	Lum
$10^{-9}$	Hyg+2Apr	-	-	-	-
$10^{-7}$	$\frac{1}{2}$ Hyg+ $\frac{1}{2}$ Apr	1	1	-	-
$10^{-8}$	$\frac{1}{2}$ Hyg+ $\frac{1}{2}$ Apr	-	-	1	1
$10^{-9}$	$\frac{1}{2}$ Hyg+ $\frac{1}{2}$ Apr	-	-	-	-
$10^{-10}$	$\frac{1}{2}$ Hyg+ $\frac{1}{2}$ Apr	-	-	-	-
$10^{-7}$	No antibiotic	-	-	2	2
$10^{-8}$	No antibiotic	2	2	5	5
$10^{-9}$	No antibiotic	7	7	5	5
$10^{-10}$	No antibiotic	29	28	-	-

Table 5.8: Excision efficiency at 48 hours. The selective media is denoted by SFM+Hyg+Apr where Hyg and Apr denote 12.5  $\mu\text{g}/\text{mL}$  of hygromycin and apramycin respectively. The 12.5  $\mu\text{g}/\text{mL}$  concentration is doubled or halved to 25  $\mu\text{g}/\text{mL}$  or 6.25  $\mu\text{g}/\text{mL}$  respectively as denoted in the table. The abbreviations Col and Lum denote the total number of colonies and the number of those colonies that exhibited luminescence respectively. Dashes denote plates on which no growth occurred and therefore could not give a light reading.

vast majority of colonies that grew exhibited luminescence. This was expected for all L3+pCC4 colonies, the L3+pRDF1 colonies grown on selective media and the L3+pRDF1 colonies grown on non-selective media with added antibiotic, but not for the L3+pRDF1 colonies grown on non-selective media without added antibiotic. Again, there is no indication that the gp3 was having an effect since the two colonies that grew on non-selective SFM plates without added antibiotic were both luminescent.

Bacterial growth occurred on just 3 of the 36 plates at the 72 hour time point

(Table 5.9), with all colonies exhibiting luminescence. All six L3+pCC4 colonies

72 hours (+72 hours)		L3+pRDF1			
		SFM		SFM+Hyg+Apr	
Dilution		Col	Lum	Col	Lum
$10^{-4}$	Hyg+2Apr	-	-	-	-
$10^{-3}$	$\frac{1}{2}$ Hyg+ $\frac{1}{2}$ Apr	2	2	-	-
$10^{-4}$	$\frac{1}{2}$ Hyg+ $\frac{1}{2}$ Apr	4	4	-	-
$10^{-5}$	$\frac{1}{2}$ Hyg+ $\frac{1}{2}$ Apr	-	-	-	-
$10^{-6}$	$\frac{1}{2}$ Hyg+ $\frac{1}{2}$ Apr	-	-	-	-
$10^{-5}$	No antibiotic	-	-	-	-
$10^{-6}$	No antibiotic	-	-	-	-
$10^{-7}$	No antibiotic	1	1	-	-
$10^{-8}$	No antibiotic	-	-	-	-

		L3+pCC4			
		SFM		SFM+Hyg+Apr	
Dilution		Col	Lum	Col	Lum
$10^{-9}$	Hyg+2Apr	-	-	-	-
$10^{-7}$	$\frac{1}{2}$ Hyg+ $\frac{1}{2}$ Apr	-	-	-	-
$10^{-8}$	$\frac{1}{2}$ Hyg+ $\frac{1}{2}$ Apr	-	-	-	-
$10^{-9}$	$\frac{1}{2}$ Hyg+ $\frac{1}{2}$ Apr	-	-	-	-
$10^{-10}$	$\frac{1}{2}$ Hyg+ $\frac{1}{2}$ Apr	-	-	-	-
$10^{-7}$	No antibiotic	-	-	-	-
$10^{-8}$	No antibiotic	-	-	-	-
$10^{-9}$	No antibiotic	-	-	-	-
$10^{-10}$	No antibiotic	-	-	-	-

Table 5.9: Excision efficiency at 72 hours. The selective media is denoted by SFM+Hyg+Apr where Hyg and Apr denote 12.5  $\mu\text{g}/\text{mL}$  of hygromycin and apramycin respectively. The 12.5  $\mu\text{g}/\text{mL}$  concentration is doubled or halved to 25  $\mu\text{g}/\text{mL}$  or 6.25  $\mu\text{g}/\text{mL}$  respectively as denoted in the table. The abbreviations Col and Lum denote the total number of colonies and the number of those colonies that exhibited luminescence respectively. Dashes denote plates on which no growth occurred and therefore could not give a light reading.

that grew at this time point on non-selective SFM plates with added antibiotic were luminescent as expected. However, the single colony that grew on non-selective SFM plates without added antibiotic was also luminescent and hence the effect of gp3 was unclear over the 72 hours of growth. An explanation for this could be that the gp3 was working efficiently, but that the integrase in the system was causing the *luxCDABE* genes to be integrated back into the *S. coelicolor* genome and thus re-establishing luminescence output.

### 5.3 Conclusions

We have carried out a series of practical experiments in order to collect primary data relating to the efficiency of DNA recombination *in vivo* using *S. coelicolor* bacteria. We initially determined that the most suitable method for growing bacterial cultures was through picking colonies from a starter plate that were seen to exhibit luminescence rather than diluting and plating directly from the glycerol stocks. We were then able to grow L3+pRDF1 cultures, containing the appropriate *luxCDABE* operon and *gp3* gene, in order to examine the efficiency of the RDF in the excision of the *luxCDABE* operon.

Growth was generally unsuccessful, with the majority of plates showing no growth. Analysis of the well developed colonies was inconclusive. The colonies we expected to show the effect of gp3-pSAM2-mediated excision were those L3+pRDF1 strains grown on selective and non-selective SFM plates. Of these, the non-selective plates represented the best medium for recording excision efficiency. A total of thirty-five such colonies grew across the four time points however, all of these colonies exhibited luminescence with the exception of four colonies at the 0 hour time point. We would have expected the number of non-luminescent colonies to increase over time as the gp3 would be able to excise more *luxCDABE* genes in conjunction with the natural production of pSAM2 integrase. In contrast, the control experiments and the L3+pRDF1 colonies grown on selective media largely produced the expected result of 100% luminescent colonies. The mechanistic model we have developed (described in Chapter 3) predicted that the excision reaction would be susceptible to low efficiency based on the duality of its mediation and thus aligns with our experimental data that suggest the gp3 was functional, but that the integrase in the system was causing natural re-integration of the *luxCDABE* operon and thus re-establishing luminescence output. The period of time during which excision has occurred, but the gene has not yet been re-integrated appears to be small and not conducive to experimental measurement. Hence, we have tangible evidence that the standard genetic switch may struggle to provide the hold states required to regulate gene expression to the desired degree.

As more time elapsed, less growth was observed with only 3 plates out of 36 exhibiting growth at 72 hours. Since the SFM plates presented a finite quantity of nutrients and moisture to the developing *S. coelicolor* colonies, which would have diminished over time, the bacteria may have died naturally before reaching our final time point. Another factor to consider is that the integration of new genes into the bacterial genome is likely to place additional burdens on cell resources

which could have a number of impacts including depleted antibiotic resistance and general functional failures that contribute to premature cell death. We decreased the concentrations of antibiotic used to select for the appropriate *S. coelicolor* strains in the hope that these stresses placed on the cells would be reduced, however the overall lack of growth we observed makes it unclear what effect this had.

We can conclude that further experimentation is required to record data worthy of elucidating the excision efficiency that we are interested in. The experimentation carried out involved  $\sim 300$  man-hours staggered over the course of 12 months. This is a substantial time frame considering that the results obtained were inconclusive and warrant further work and, hence, we have demonstrated the technical difficulty in achieving sound experimental results as well as the time-consuming nature of the necessary procedures. As the complexity of synthetic circuitry increases, the relevant experimental procedures will undoubtedly require even greater technical knowledge and time frames to achieve successful implementation. The development of sophisticated mechanistic models that are capable of quantitative dynamical predictions will therefore be invaluable in realising the practical application of these novel systems.

## 5.4 Methods

### 5.4.1 Intergenic conjugation protocol

The transfer of vectors into *S. coelicolor* M145 was carried out using the protocol specified in Kieser et al. [2000]. Single colonies of ET12567 cells with pUZ8002 containing the relevant vector were picked and grown overnight at 37 °C shaking in lysogeny broth (LB) with the appropriate antibiotics. The next morning 200  $\mu\text{L}$  of this starter culture was used to inoculate 10 mL fresh media (with the same antibiotics) and this was grown at 37 °C shaking until the  $\text{OD}_{600}$  was between 0.4 and 0.6 ( $\sim 4$  hours). This was then centrifuged for 10 minutes at 2000 rpm to pellet the cells. The pellet was then re-suspended in 10 mL LB and centrifuged again before the washing step was repeated to remove any remaining antibiotics. The cell pellet was then re-suspended in the residual LB to give a total volume of 1 mL.

A volume of 10  $\mu\text{L}$  *Streptomyces* spore stock was added to 500  $\mu\text{L}$  2xYT media and the cells heat-shocked at 50 °C for 10 minutes before being mixed with 500  $\mu\text{L}$  of the prepared ET12567 cells. This mixture was then serially diluted and the two strains were grown overnight together on SFM media on four different plates containing dilutions of between  $10^{-1}$  and  $10^{-4}$ . The next morning the plates were overlaid with nalidixic acid to kill the *E. coli* and apramycin or hygromycin to select

for *Streptomyces* colonies contain the luciferase constructs or pCC4 vectors. This was then left to grow for 3 to 4 days, when single colonies could be collected and used to inoculate fresh plates.

#### 5.4.2 Media stock solutions

##### **SFM**

8 g bacto-agar

8 g soya flour

8 g mannitol

Make up to 400 mL with tap water  
and mix together before autoclaving

##### **2xYT**

16 g tryptone

10 g yeast extract

5 g NaCl

Make up to 1 L with distilled water,  
adjust the pH to 7.0 and mix together  
before autoclaving

Autoclaving was carried out at 121°C for 20 minutes, with media then stored at room temperature. Once antibiotics were added the media would be used immediately or stored in the fridge until required [Styles, 2016].

## Chapter 6

# Mechanistic Modelling of the Regulatory System Controlling Methylenomycin Production in *Streptomyces coelicolor*

### 6.1 Scientific background

#### 6.1.1 A brief history of antibiotics

The first known antibiotic was famously discovered by mistake by Alexander Fleming in 1928. After returning from holiday, Fleming noticed that mould had formed on his discarded Petri dishes containing cultures of the bacteria *Staphylococcus* (Figure 6.1). He could see that there were areas around the mould where the bacteria had been killed and, after further experimentation, he identified the active agent penicillin. It was 1942 when penicillin was eventually cultivated sufficiently to treat bacterial infections, a milestone that has revolutionised medicine as we know it through the development of a variety of antibiotics enabling wide-ranging treatment of numerous bacterial infections including pneumonia, tuberculosis and bacterial meningitis. That said, an overwhelming eradication of many of the most pathogenic infections that might have been predicted has not been realised due to the phenomenon of bacterial resistance. Resistance to antibiotics was observed by Fleming even before penicillin production reached a commercial scale, and has subsequently threatened to nullify the therapeutic effects of any currently known antibiotics [Brown, 2005].

In the early 1950s, erythromycin was introduced as an alternative to penicillin

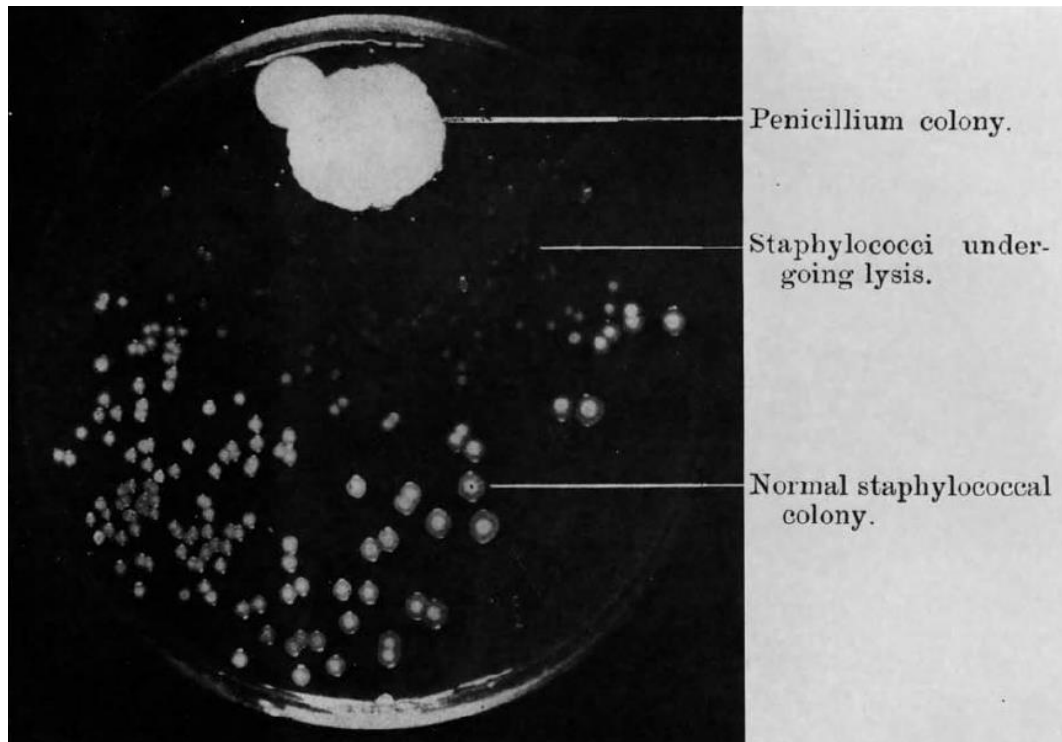


Figure 6.1: Photograph showing the antibacterial effect of a penicillium colony on staphylococcal colonies, taken from Fleming [1980].

at Boston City Hospital in an attempt to combat *Staphylococcus aureus* infections. Unfortunately, in less than a year, approximately 70% of *S. aureus* infections were demonstrating resistance to erythromycin and the antibiotic was completely withdrawn as a result [Finland, 1979]. A number of factors are thought to contribute to the cause of such significant biological adaptation; “*there is perhaps no better example of the Darwinian notions of selection and survival*” [Davies and Davies, 2010]. Human activities have contributed significantly to the prevalence of bacterial resistance. Underuse of antibiotics allows target bacteria that survive an insufficient dose to proliferate in the absence of individuals with the greater susceptibility to the antibiotic, creating a population with decreased susceptibility. Furthermore, overuse and misuse of antibiotics in situations where they are not required can cause otherwise harmless bacteria that encounter the antibiotics to develop resistance and become pathogenic.

The mechanisms that facilitate the development of bacterial resistance include the aforementioned natural selection that enables bacteria with certain genetic predispositions to evolve into populations possessing highly effective resistance to antibiotics. Bacteria are also capable of transferring genetic information between

individuals, which enables resistance genes to spread throughout populations. This is referred to as horizontal gene transfer (HGT) and is considered the primary cause of bacterial resistance. Resistance can also be caused by random mutations in the bacterial DNA, in the event of which the bacteria would, again, possess an evolutionary advantage that could easily proliferate quickly throughout populations be it through selective pressures or intercellular transmission.

As a result of approximately 75 years of antibiotic misuse, so called ‘superbugs’ have emerged, some that are multidrug resistant (MDR), some that are extremely drug resistant (XDR), that is, resistant to at least four foremost associated treatments, and others that appear to be resistant to all available antimicrobial treatments, or totally drug resistant (TDR). The use of the antibiotics streptomycin and isoniazid initially demonstrated considerable success in the treatment of *Mycobacterium tuberculosis* infections however the rapid development of resistance, thought to arise exclusively by virtue of spontaneous mutation [Davies and Davies, 2010], has led to the emergence of XDR strains of this bacteria which infect approximately one-third of the entire world population, in both developed and developing nations. Arguably the most publicised example of a superbug is methicillin-resistant *S. aureus* (MRSA) which, despite lacking the historical notoriety of *M. tuberculosis*, has become the primary cause of bacterial infection in hospital environments. Again, the initial treatment of *S. aureus* infections using penicillin and the subsequent shift towards the newly derived methicillin was thought to inhibit the action of this and other similar bacteria. However, within 3 years mutant strains appeared demonstrating resistance to several recommended antibiotics and hence, today, the MRSA acronym is used to describe multidrug-resistant *S. aureus*.

It is clear that the magnitude of the implications related to antibiotic resistance cannot be overestimated, with many experts predicting a return to the days before Fleming’s revolutionary discovery if a new global approach is not adopted urgently. Systems biology approaches have improved understanding of the mechanisms associated with the action of antibiotics on target bacteria and the propagation of resistance in these microbes [Yeh et al., 2009]. One particular mathematical modelling investigation yields direct predictions regarding the impact of drug synergy on the emergence of resistance [Michel et al., 2008]. The model is capable of predicting the effects of different multidrug combinations on the development of bacterial resistance, with results suggesting that synergistic multidrug combinations are conducive to the development of resistance, whereas antagonistic multidrug combinations are actually conducive to the inhibition of resistance. Synergistic multidrug combinations are typically favoured by clinicians due to the increased efficacy that arises



from a broader spectrum of activity however, it appears that greater consideration of the trade-off between efficacy and resistance limitation is required.

The study of soil-dwelling bacteria is integral to combating antibiotic resistance. These microorganisms are not only responsible for antibiotic production, but are also targeted by antibiotics produced by other bacterial populations. Synthetic biology approaches place the native systems known to mediate antibiotic production at the forefront of new research efforts in order to elucidate fundamental mechanistic properties. Specialised gene clusters have been identified in model bacterial strains that are known to regulate the production of useful antibiotics [Zou et al., 2017], however there remains uncertainty regarding their structural composition and the roles and actions of the associated regulatory elements. Examination of regulatory gene clusters will identify highly functional motifs, the characterisation of which through computational analyses will inform the assembly of novel synthetic antibiotic production circuits. Hence, mathematical modelling approaches will be key to the systems-level elucidation of antibiotic action that can assist in delivering measures devised to prevent resistance, as well as facilitating the discovery of new antibiotics.

### 6.1.2 *Streptomyces*

*Streptomyces* refers to the genus of streptomycetes that represent the largest family of the actinomycetes (actinobacteria) phylum [Flärdh and Buttner, 2009]. These Gram-positive bacteria naturally inhabit soil, producing the natural product geosmin that is known to give the soil its typically earthy smell [Gerber and Lechevalier, 1965]. Despite being bacteria, *Streptomyces* exhibit behaviour that resembles fungi, facilitating the decomposition of dead organisms and vegetation that contributes to the nitrogen cycle. They bridge the gap between bacteria and fungi further by virtue of their complex life cycle. *Streptomyces* begin life as a spore that germinates in the presence of the appropriate nutrients. This causes the formation of vegetative hyphae that extend into the immediate environment in becoming fungi-like mycelia. Non-branching sporogenic aerial hyphae are formed in the event of nutrient depletion [Flärdh and Buttner, 2009]. The life cycle is restarted as these aerial hyphae partition to form largely dormant unigenomic spores. During this spore formation stage, a wide range of secondary metabolites and natural products are synthesised [Jakimowicz and van Wezel, 2012]. Streptomycetes produce ~70% of all commercial antibiotics currently available, thus demonstrating their immense importance in the discovery of new natural products [Watve et al., 2001].

### 6.1.3 The methylenomycin regulatory gene cluster

The bacterium *Streptomyces coelicolor* A3(2) has emerged as the model organism for studying streptomyces, initially thanks to the production of coloured metabolites that facilitated genetic studies, and more recently thanks to the sequencing of its entire genome [Bentley et al., 2002]. These bacteria have a 8,667,507 base pair single linear chromosome containing protein coding genes of which over 12% are thought to be regulatory [Bentley et al., 2002]. These predicted transcriptional regulators are thought to mediate antibiotic synthesis through the production of microbial hormones, as well as influence structural and metabolic cellular responses [Willey and Gaskell, 2011]. The linear SCP1 plasmid (~356 kb) and the circular SCP2 plasmid (~31 kb) are both present within the *S. coelicolor* genome and have also both been sequenced [Bentley et al., 2004]. This genome sequencing has revealed many cryptic and ‘silent’ gene clusters: sets of genes predicted to produce a natural product, but whose product has not been observed. Silent gene clusters have been awakened through genetic manipulation of regulatory elements [Sidda et al., 2014; Laureti et al., 2011]. Thus, characterisation of the regulatory system that mediates the production of specialised metabolites is key to discovering new natural products. Developing improved understanding of the regulatory architectures that underlie natural product biosynthesis can also accelerate the design of novel regulatory systems in synthetic biology.

The antibiotic methylenomycin A is a natural product of *S. coelicolor* A3(2) and is of particular interest since all of the 21 biosynthetic, regulatory and resistance genes, located in a cluster on the SCP1 plasmid [Bentley et al., 2002], have been studied in detail [Corre and Challis, 2005], and a series of knockout mutant strains has been generated [O’Rourke et al., 2009]. The regulation of methylenomycin biosynthesis is mediated by the transcriptional repressor MmfR, a TetR-family homodimeric protein consisting of an N-terminal DNA-binding domain and a C-terminal ligand-binding domain (Fig. 6.2A) [Ramos et al., 2005; Corre, 2013]. In the initial growth phase of *S. coelicolor*, the MmfR N-terminal domain is thought to be bound to the DNA at the methylenomycin auto-regulatory response element (MARE) causing the transcriptional repression of downstream genes. MmfR holds the system in this repressed state until the advent of the small signalling molecules, methylenomycin furans (MMFs) [Corre et al., 2008]. MMFs bind specifically to the C-terminal domain of the MmfR, forming an MmfR:MMF complex that results in a conformational change in the MmfR. Consequently, MmfR is released from the MARE, negating the repression and triggering gene transcription. The biosynthesis of MMFs is controlled by the MmfLHP enzymes which are, themselves, repressed by

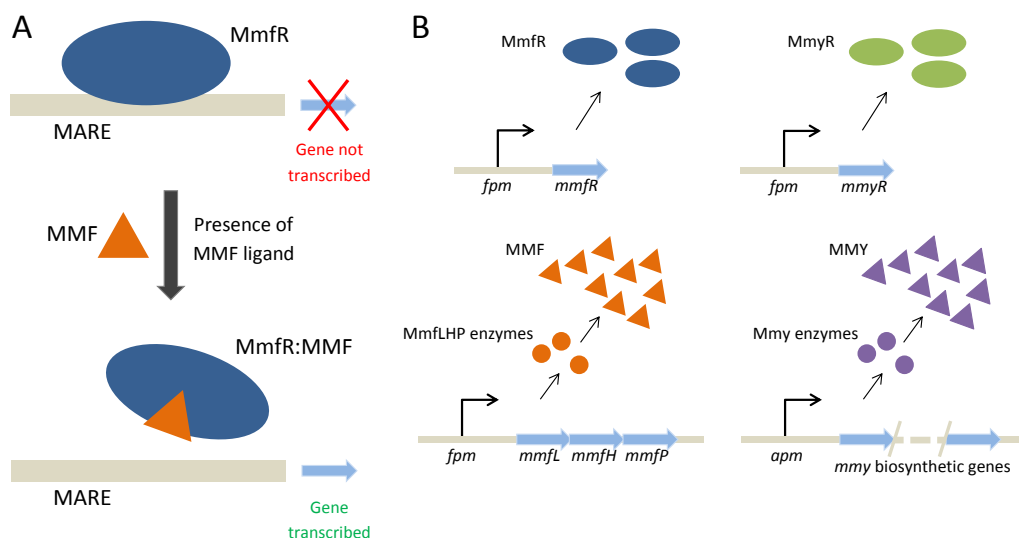


Figure 6.2: A) Schematic diagram of the MmfR binding mechanism. Binding of MmfR to DNA at the MARE represses gene transcription and therefore negates system output. In the presence of MMF ligand, an MmfR:MMF complex is formed which releases MmfR from the MARE and triggers gene transcription. B) Schematic diagram of the methylenomycin gene cluster whereby *fpm* and *apm* represent the DNA binding motifs recognised by MmfR and MmyR proteins. The *fpm* controls the expression of *mmfR*, *mmyR* and *mmfLHP* genes while *apm* regulates the expression of the *mmy* biosynthetic genes.

MmfR, thus forming a feedback control loop that governs the dynamical properties of the system. A second repressor, MmyR, is homologous to MmfR yet its role in methylenomycin regulation is currently less understood. There is, however, clear evidence that the impact of MmyR is particularly significant, since *S. coelicolor* strains with the *mmyR* gene knocked out have been found to over-produce methylenomycin [Chater and Bruton, 1985].

Homologous architectures to that of the methylenomycin regulatory system have been identified across a plethora of microorganisms [Liu et al., 2013], regulating different classes of natural products and thus indicating the utility of this specific type of regulatory architecture [Corre et al., 2008]. Responding to environmental changes is of paramount importance to these bacteria. The soil they live in presents a harsh environment with considerable competition for resources and it is therefore vital that they possess sophisticated, tightly regulated mechanisms to turn on the expression of specific genes when required. Hence, obtaining a detailed mechanistic understanding of the regulatory system controlling the biosynthetic pathway to this antibiotic has the potential to elucidate a host of other, less tractable, biosynthetic

gene clusters and help standardise one of the most important regulatory networks for the development of new antibiotics.

## 6.2 Formulation of candidate model architectures

The various binding interactions and protein expression summarised in Fig. 6.2 inform the formulation of our candidate model architectures. MmfR is thought to bind to three distinct intergenic regions on the gene cluster [O’Rourke et al., 2009]. However, we combine the region associated with MmyR biosynthesis together with the region associated with both MmfR and MMF biosynthesis to form a single DNA module responsible for the biosynthesis of all three molecules (the furan producing module, *fpm*). That is, we use the term *fpm* to refer to five distinct genes that provide control over three distinct molecular products: MmyR, MmfR and MMF. The genes *mmfL*, *mmfH* and *mmfP* are coregulated in an operon and are directly responsible for the production (assembly) of MMF molecules; the *mmfR* and *mmyR* genes control MmfR and MmyR production respectively [Corre et al., 2008; O’Rourke et al., 2009] (Fig. 6.2B). The third distinct intergenic region is represented by our second DNA module which we consider responsible for methylenomycin (MMY) biosynthesis only (the antibiotic producing module, *apm*). Therefore, our model architectures all consist of two fundamental DNA modules that can both be bound by MmfR, and that have production of their respective proteins repressed as a consequence. Due to its effect on the gene cluster and its homology to MmfR, we assume that MmyR also binds both modules in a similar manner.

Our base architecture accounts for reversible MmfR and MmyR binding to both the *fpm* and *apm* to form four complexes: *fpm*:MmfR, *fpm*:MmyR, *apm*:MmfR and *apm*:MmyR. MMF binds MmfR reversibly at these complexes in order to trigger gene expression; MMF binding MmfR in solution is also accounted for since we have been able to co-crystallise MmfR:MMF complexes and solve the 3D-structure through experimentation void of target DNA modules (data not shown). MmfR:MMF complexes that dissociate from the MAREs return free MmfR and MMF back into the system irreversibly. MmyR, MmfR and MMF production is controlled by the *fpm*. We account for an initial repressed system state by imposing non-zero initial concentrations upon the *fpm*:MmfR and *apm*:MmfR complexes; all remaining model variables have initial concentrations equal to 0. MmfR, MmyR, MMF and MMY all undergo degradation at constant rates (Fig. 6.3).

This model architecture represents the extent of our current mechanistic understanding, however there are certain details that require further investigation. For

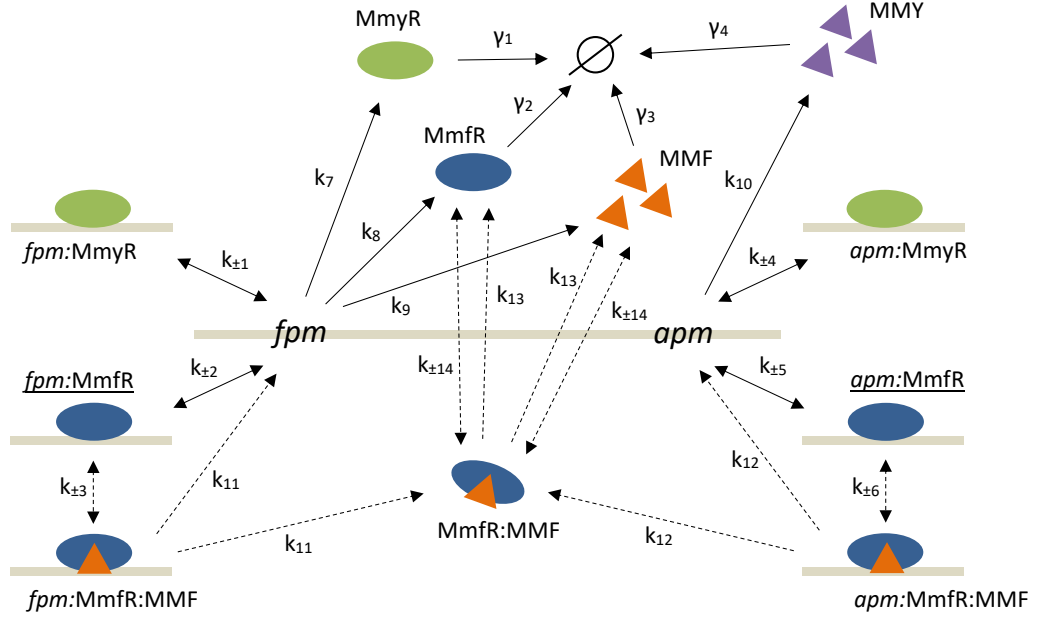


Figure 6.3: Schematic diagram of the reaction network comprising the base (BNN) model architecture. Reversible and irreversible reactions are depicted by double and single arrows respectively; reaction rate constants are denoted by the corresponding numbered  $k$ . The empty set depicts protein degradation, with rate constants denoted by the corresponding numbered  $\gamma$ . Solid arrows depict reactions that are common to all 48 model architectures, whereas dashed arrows depict those that are subject to adaptation. Cellular entities with non-zero initial concentrations are underlined.

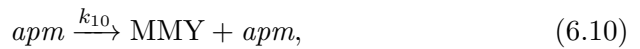
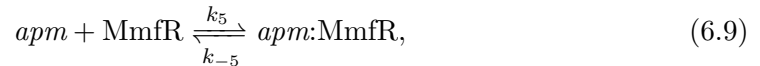
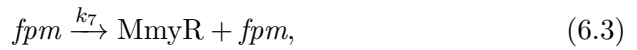
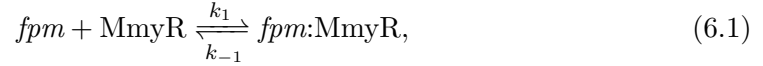
example, although we believe that the MMF releases MmfR from the *fpm*:MmfR and *apm*:MmfR complexes and also binds free MmfR in solution, it would be insightful to examine the dynamical influence of each binding mechanism in isolation. Similarly, although we believe there is no interaction between the MMF and MmyR within the system (data not shown), the binding interactions of MMY are not as well documented. It may therefore be possible that MMY is able to inhibit the action of both MmfR and MmyR either through dissociation from their respective *fpm* and *apm* complexes or binding in solution. Consequently, the aim of our modelling investigation is to examine the effect of three key mechanistic properties on model performance:

1. MMF-MmfR interactions occur at existing DNA:MmfR complexes (C), in solution (S) or via both mechanisms (B).
2. MMY-MmfR interactions occur at existing DNA:MmfR complexes (C), in solution (S), via both mechanisms (B) or do not occur at all (N).

3. MMY-MmyR interactions occur at existing DNA:MmyR complexes (C), in solution (S), via both mechanisms (B) or do not occur at all (N).

This set of possible molecular interactions results in 48 distinct candidate model architectures for the methylenomycin regulatory system. Each candidate architecture is given a three letter name corresponding to the interactions accounted for with respect to each of the three properties listed above. The order of the letters in each name corresponds strictly to the numerical order of these properties. For example, our base architecture (described above) is given the name BNN since it accounts for both mechanisms (B) regarding property 1, for no interactions at all (N) regarding property 2 and for no interactions at all (N) regarding property 3.

Each of the 48 candidate architectures presents a distinct reaction network and set of biochemical equations that can be used to derive a dynamical mathematical model. We apply mass action kinetics to the biochemical equations comprising each reaction network to derive a corresponding system of ordinary differential equations (ODEs). Each ODE describes the rate of change in concentration corresponding to each model variable (cellular entity). The solution to each system of ODEs is determined numerically due to the non-linearity of the equations and provides a deterministic output that can be used to simulate and predict *in vivo* system dynamics *in silico*. For example, the BNN model architecture is comprised of the following biochemical equations:



$$apm:MmfR + MMF \xrightleftharpoons[k_{-6}]{k_6} apm:MmfR:MMF, \quad (6.11)$$

$$apm:MmfR:MMF \xrightarrow{k_{12}} apm + MmfR:MMF, \quad (6.12)$$

$$MmfR:MMF \xrightarrow{k_{13}} MmfR + MMF, \quad (6.13)$$

$$MmfR + MMF \xrightleftharpoons[k_{-14}]{k_{14}} MmfR:MMF, \quad (6.14)$$

$$MmyR \xrightarrow{\gamma_1} \emptyset, \quad (6.15)$$

$$MmfR \xrightarrow{\gamma_2} \emptyset, \quad (6.16)$$

$$MMF \xrightarrow{\gamma_3} \emptyset, \quad (6.17)$$

$$MMY \xrightarrow{\gamma_4} \emptyset, \quad (6.18)$$

from which we derive the following system of model ODEs:

$$\begin{aligned} \frac{d[MmyR]}{dt} = & k_7[fpM] - k_1[MmyR][fpM] + k_{-1}[fpM:MmyR] - k_4[MmyR][apM] + \dots \\ & k_{-4}[apM:MmyR] - \gamma_1[MmyR], \end{aligned} \quad (6.19)$$

$$\begin{aligned} \frac{d[MmfR]}{dt} = & k_8[fpM] - k_2[MmfR][fpM] + k_{-2}[fpM:MmfR] - k_5[MmfR][apM] + \dots \\ & k_{-5}[apM:MmfR] + k_{11}[fpM:MmfR:MMF] + k_{12}[apM:MmfR:MMF] + \dots \\ & k_{13}[MmfR:MMF] - k_{14}[MmfR][MMF] + k_{-14}[MmfR:MMF] - \gamma_2[MmfR], \end{aligned} \quad (6.20)$$

$$\begin{aligned} \frac{d[fpM]}{dt} = & k_{11}[fpM:MmfR:MMF] - k_1[MmyR][fpM] + k_{-1}[fpM:MmyR] + \dots \\ & k_{-2}[fpM:MmfR] - k_2[MmfR][fpM], \end{aligned} \quad (6.21)$$

$$\begin{aligned} \frac{d[apM]}{dt} = & k_{12}[apM:MmfR:MMF] - k_4[MmyR][apM] + k_{-4}[apM:MmyR] + \dots \\ & k_{-5}[apM:MmfR] - k_5[MmfR][apM], \end{aligned} \quad (6.22)$$

$$\frac{d[fpM:MmyR]}{dt} = k_1[MmyR][fpM] - k_{-1}[fpM:MmyR], \quad (6.23)$$

$$\frac{d[apM:MmyR]}{dt} = k_4[MmyR][apM] - k_{-4}[apM:MmyR], \quad (6.24)$$

$$\begin{aligned} \frac{d[fpM:MmfR]}{dt} = & k_2[MmfR][fpM] - k_{-2}[fpM:MmfR] - k_3[fpM:MmfR][MMF] + \dots \\ & k_{-3}[fpM:MmfR:MMF], \end{aligned} \quad (6.25)$$

$$\begin{aligned} \frac{d[apM:MmfR]}{dt} = & k_5[MmfR][apM] - k_{-5}[apM:MmfR] - k_6[apM:MmfR][MMF] + \dots \\ & k_{-6}[apM:MmfR:MMF], \end{aligned} \quad (6.26)$$

$$\frac{d[fpM:MmfR:MMF]}{dt} = k_3[fpM:MmfR][MMF] - k_{-3}[fpM:MmfR:MMF] - k_{11}[fpM:MmfR:MMF], \quad (6.27)$$

$$\frac{d[apM:MmfR:MMF]}{dt} = k_6[apM:MmfR][MMF] - k_{-6}[apM:MmfR:MMF] - k_{12}[apM:MmfR:MMF], \quad (6.28)$$

$$\begin{aligned} \frac{d[MMF]}{dt} = & k_9[fpM] - k_3[fpM:MmfR][MMF] + k_{-3}[fpM:MmfR:MMF] + \dots \\ & k_{-6}[apM:MmfR:MMF] - k_6[apM:MmfR][MMF] + k_{11}[fpM:MmfR:MMF] + \dots \\ & k_{12}[apM:MmfR:MMF] + k_{13}[MmfR:MMF] - k_{14}[MmfR][MMF] + \dots \\ & k_{-14}[MmfR:MMF] - \gamma_3[MMF], \end{aligned} \quad (6.29)$$

$$\frac{d[\text{MMY}]}{dt} = k_{10}[\text{apm}] - \gamma_4[\text{MMY}], \quad (6.30)$$

$$\begin{aligned} \frac{d[\text{MmfR:MMF}]}{dt} = & k_{11}[\text{fpm:MmfR:MMF}] + k_{12}[\text{apm:MmfR:MMF}] - k_{13}[\text{MmfR:MMF}] + \dots \\ & k_{14}[\text{MmfR}][\text{MMF}] - k_{-14}[\text{MmfR:MMF}], \end{aligned} \quad (6.31)$$

where square brackets denote concentration and the reaction rate constants translate to model parameters, denoted by each numbered  $k$ . Reactions associated with reversible DNA:protein binding ( $k_1, k_{-1}, k_2, k_{-2}, k_4, k_{-4}, k_5$  and  $k_{-5}$ ), the production of MmyR, MmfR, MMF and MMY ( $k_7, k_8, k_9$  and  $k_{10}$ ) and each individual protein degradation reaction ( $\gamma_{1,2,3,4}$ ) are common to all of our candidate model architectures. Other reactions that are associated with the release of MmfR from existing DNA:MmfR complexes or the sequestration of MmfR and MmyR via binding in solution are not common to all models and are thus subject to investigation through our computational simulations.

Model simulations are provided by the numerical solutions to the relevant model ODEs, which are calculated using the ODE solver ode45 in MATLAB. We are interested in examining the dynamics of methylenomycin production in each of the 48 candidate models and therefore analyse the simulations of MMY provided by numerical solutions to the corresponding ODE (6.30).

### 6.3 Available experimental data

Methylenomycin production by *S. coelicolor* has been shown to adopt a typical dynamical profile [Hobbs et al., 1992; Hayes et al., 1997]. Once expression is initiated, usually by environmental conditions that are thought to establish MMF production, it increases relatively quickly towards a global maximum level. Expression then decreases from this maximum, reaching a relatively low level at steady-state. This profile aligns with the premise that the system is initially held in a repressed state until MmfR is released by MMF to trigger methylenomycin expression, which then increases quickly until free MmfR and MmyR cause secondary repression and eventual equilibrium of the feedback loop.

We consider the binding affinity of MmfR to the *fpm* and *apm* to be strong, based on experimental data regarding binding interactions between a similar protein, SAV2270, and its associated DNA motifs (our unpublished data). We characterised the binding of this protein to Streptavidin Immobilized oligonucleotides using a Biocore T200 SPR instrument. Our data reveal that the association and dissociation rates of this protein:DNA binding are on the order of  $10^5 \text{ M}^{-1}\text{s}^{-1}$  and  $10^{-2} \text{ s}^{-1}$  respectively. As a result, we fix the model parameters relating to MmfR associ-



ation and dissociation from both the *fp*m and *ap*m at  $10^5$  and  $10^{-2}$  respectively ( $k_2 = k_5 = 10^5$ ;  $k_{-2} = k_{-5} = 10^{-2}$ ). The dimensionality of our experimental measurements agree with the corresponding parameters in our dimensional model and we are therefore able to apply these values directly. We assume that MmyR binding interactions are identical to that of MmfR and hence the same values are fixed for the parameters describing MmyR association and dissociation from the *fp*m and *ap*m ( $k_1 = k_4 = 10^5$ ;  $k_{-1} = k_{-4} = 10^{-2}$ ).

Mutant strains of *S. coelicolor* that account for specific gene knockouts reveal qualitatively different methylenomycin production dynamics (Table 6.1). The

<i>S. coelicolor</i> strain	Methylenomycin production
Wildtype	+
$\Delta mmyR$	+++
$\Delta mmfLHP$	-
$\Delta mmfLHP + \Delta mmyR + \Delta mmfR$	+++
$\Delta mmfLHP + \text{exogenous MMF}$	+

Table 6.1: The effects of knocking out certain genes and combinations of genes observed experimentally, adapted from O’Rourke et al. [2009]. The wildtype strain is allocated a single ‘+’ to denote typical methylenomycin expression. Over-expression and the cessation of expression are denoted by ‘+++’ and ‘-’ respectively.

mutant strain accounting for *mmyR* deletion,  $\Delta mmyR$ , has been shown to exhibit increased methylenomycin expression compared to the wildtype; in the absence of MmyR, the overall capacity of the system to repress methylenomycin production is reduced and therefore the production of the antibiotic is increased. The  $\Delta mmfLHP$  strain exhibits a complete cessation of methylenomycin expression; in the absence of the *mmfLHP* genes, the system is locked in the *ap*m:MmfR complex since the expression of MmfR, MmyR and particularly MMF is prevented and thus the bound MmfR cannot be released. The  $\Delta mmfLHP + \Delta mmyR + \Delta mmfR$  strain exhibits increased methylenomycin production compared to the wildtype; in the absence of MmfR and MmyR, both initially and as a result of any subsequent production by the *fp*m, the *ap*m is able to produce methylenomycin in an unrestricted manner. The  $\Delta mmfLHP$  strain with exogenous MMF exhibits relatively similar methylenomycin expression to that of the wildtype; in the absence of endogenous MMFs, exogenous MMF permits the release of MmfR and, in turn, methylenomycin expression. Experimentation with the  $\Delta mmfR$  strain has thus far yielded inconclusive results and, as such, presents the opportunity for mathematical modelling simulations to inform future experimental studies.

## 6.4 Model selection via approximate Bayesian computation

In order to assess the potential of the 48 candidate architectures to reproduce the known characteristics of the system, we perform model selection based on approximate Bayesian computation using the ABC-SysBio software package. The procedure determines the model, from a set of candidate models, that is most likely to have produced the associated experimental data. Extensive quantitative data regarding methylenomycin expression is lacking in the literature, however a time course expression profile is reported in Hobbs et al. [1992]. We therefore provide ABC-SysBio with a dataset designed to replicate this profile (Fig. 6.4), with two important excep-

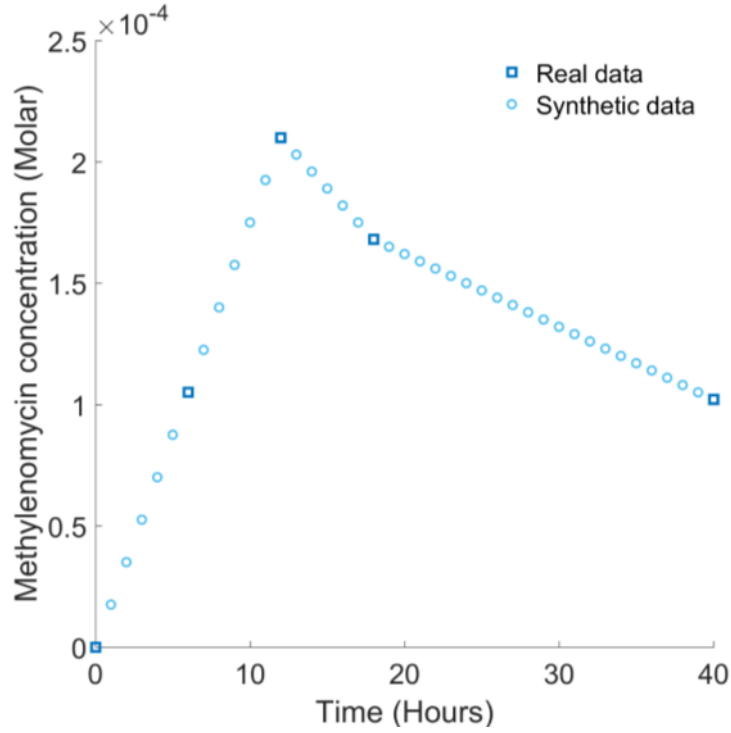


Figure 6.4: Experimental data representing current biological knowledge of typical methylenomycin expression in *S. coelicolor*. Real experimental data points taken from Hobbs et al. [1992]. Synthetic data points are added uniformly between real data points to increase the rigour of model selection and parameter inference.

tions. Firstly, we specifically account for the dynamical series of data points in the 40 hour interval between hours 54 and 94 of the time course. This is because the 54 hour experimental time point is when methylenomycin expression commences and translates to the 0 hour time point in our simulations. The time points that precede

54 hours record the repression of methylenomycin production prior to the environmental trigger and are hence excluded when fitting a model that accounts purely for the dynamical response of the system. Secondly, we incorporate additional uniformly distributed ‘synthetic’ data points, increasing the size of the dataset from 5 points to 41, in order to provide a more rigorous data fitting task to the ABC-SysBio algorithm.

ABC-SysBio also requires a prior probability distribution on each model parameter subject to inference in order to establish the parameter space within which to locate acceptable parameter sets. The prior distributions chosen for all parameters associated with each of the 48 candidate models are uniform distributions on the interval  $[10^{-4}, 10^4]$ , that is, all candidate models are given an equal parameter space in attempting to identify parameter values capable of replicating our experimental dataset. We also impose prior distributions on the initial conditions of the necessary state variables due to the lack of experimental data regarding the physical quantity of DNA in the system: the prior distributions are uniform distributions on the interval  $[0, 1]$  and are assigned only to the MmfR:*fpm* and MmfR:*apm* complexes, all other initial conditions are set equal to 0. ABC-SysBio convergence is dependent on the sequential satisfaction of a predefined series of decreasing error thresholds by a predefined number of solutions. Here, the number of solutions required to satisfy each error threshold is 500 [Woods and Barnes, 2016] in order to reduce the time frame required for convergence; the number of models subject to selection coupled with the inability to parallelise the process presents a particularly time consuming computational workload. The user-defined error function designed to measure the accuracy of simulations takes the mean absolute value of the difference between model outputs and data values:

$$E = \frac{1}{41} \sum_{i=1}^{41} |x_i - d_i|, \quad (6.32)$$

where  $E$  is the error and  $x_i$ ,  $d_i$  are the model outputs and data values at each of the 41 corresponding time points,  $t_i$ , respectively.

The results of our model selection are shown in Fig. 6.5. The final probability distributions reveal that the model most likely to have produced the experimental data is BNN, the model formulated based on our current knowledge (Fig. 6.5A). The BNN model achieved a 0.916 probability of producing our data which is vastly superior to the remaining models, 36 of which were statistically eliminated through the selection process. This suggests that the most plausible network of molecular interactions underlying this system should account for MMF-MmfR interactions both

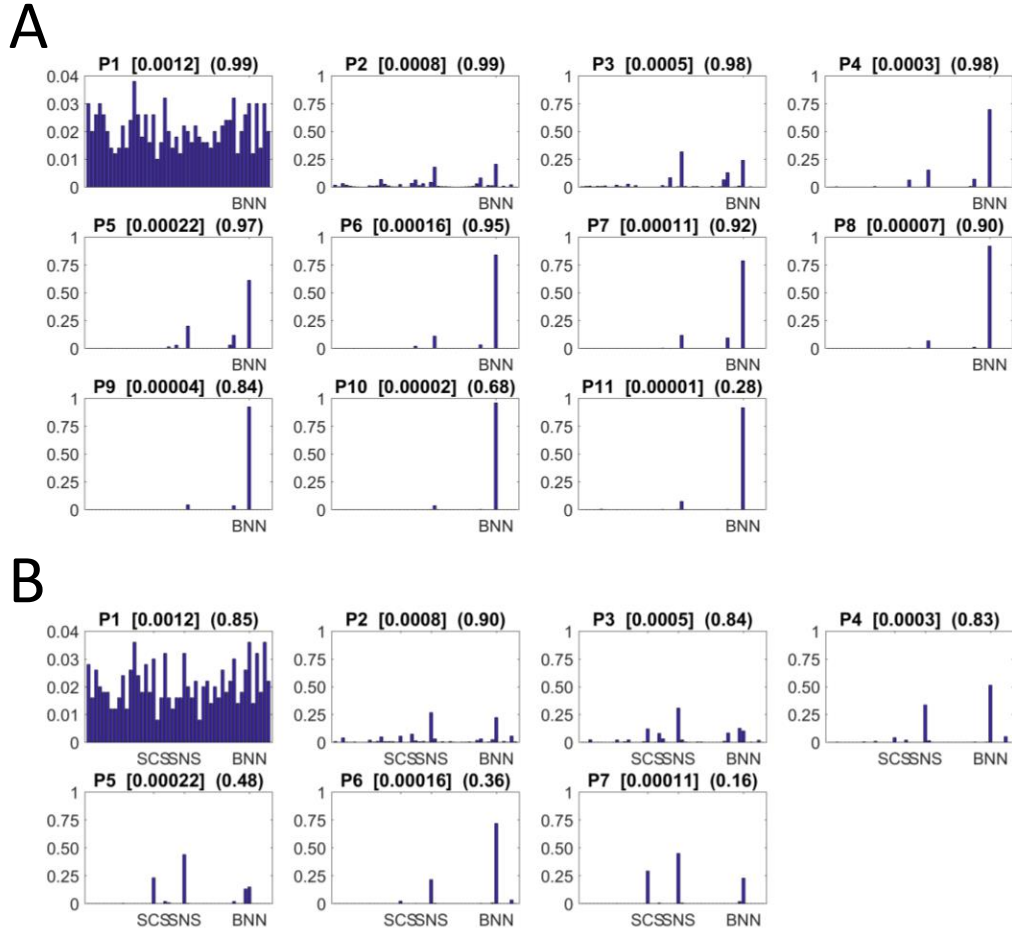


Figure 6.5: ABC-SysBio model selection results. A) Histograms showing the probabilities of producing the full dataset for the 48 candidate models. B) Histograms showing the probabilities of producing the real experimental dataset for the 48 candidate models. The numbers above each histogram denote the population number, the error threshold  $\epsilon$  (square brackets) and the acceptance rate (parentheses) respectively. The number of accepted solutions required to satisfy each error threshold is 500.

at existing DNA:MmfR complexes and in solution, no MMY-MmfR interactions at all and no MMY-MmyR interactions at all, as depicted in Fig. 6.3.

In order to verify that the addition of synthetic data points does not restrict the emergence of other viable candidate models, we repeated the model selection procedure using only the 5 real experimental data points taken from Hobbs et al. [1992]. Mean absolute error generally increases with decreasing numbers of data points which subsequently increases the difficulty for each population of solutions

to meet the same error thresholds. Hence, the acceptance rate decreases and the process becomes more time consuming; this run took longer than the original run and met 7 thresholds compared to the previous 11 (Fig. 6.5B). The probability distribution across all models clearly identified the most likely models as early as P2, which converged further at P4 and P6 to suggest that BNN was a likely model architecture, in agreement with our initial result. However, P3, P5 and P7 identified a different distribution which suggested that models SCS and SNS were also likely candidates. Given that ABC-SysBio appeared to present two alternating probability distributions, it is probable that additional local minima were identified in this case. To further investigate the set of plausible models identified using this Bayesian inference framework, we next employed global optimisation methods as well as analysis of mutant versions of the candidate models, as described in the following sections.

## 6.5 Parameter inference via global optimisation

ABC-SysBio performs parameter inference by producing probability distributions on the numerical values that comprise accepted parameter sets during the model selection process. For example, the distributions on the initial conditions imposed on the *fpm:MmfR* and *apm:MmfR* complexes reveal that statistically these values can be approximated to be 0.6 and 0.5 respectively (Fig. 6.6). These distributions

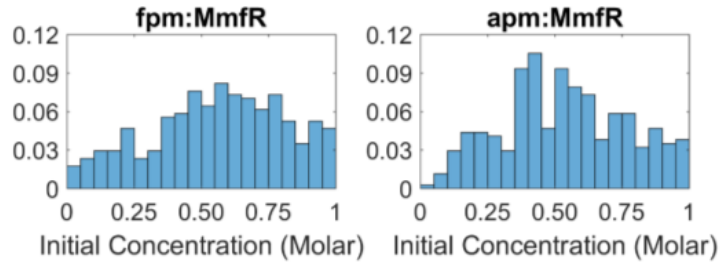


Figure 6.6: ABC-SysBio parameter inference results. Histograms show the probability distributions on the two parameters describing the initial concentration of the *fpm:MmfR* and *apm:MmfR* complexes.

are insightful, but cannot provide complete clarity over the numerical values inferred in all cases. Other parameter inference methods, such as global optimisation, place greater focus on the identification of specific numerical parameter sets capable of minimising the user-defined error function. We therefore employ the genetic algorithm (GA), a well established global optimisation method, to assess the data fitting

qualities of our BNN model.

In this case, the error function minimised by the GA is the same absolute mean error function used for ABC-SysBio model selection (6.32). We also allocate the same parameter space to the GA by imposing lower and upper bounds on the inferred parameters of  $10^{-4}$  and  $10^4$  respectively. Again, the initial conditions imposed on the model variables are 0 with the exception of those regarding the *fpm:MmfR* and *apm:MmfR* complexes which we approximate to be 0.55 in light of our ABC-SysBio probability distributions and given that we require both initial concentrations to be equal. The results of our global optimisation are shown in Fig. 6.7. The BNN model is capable of accurately matching the experimental time course

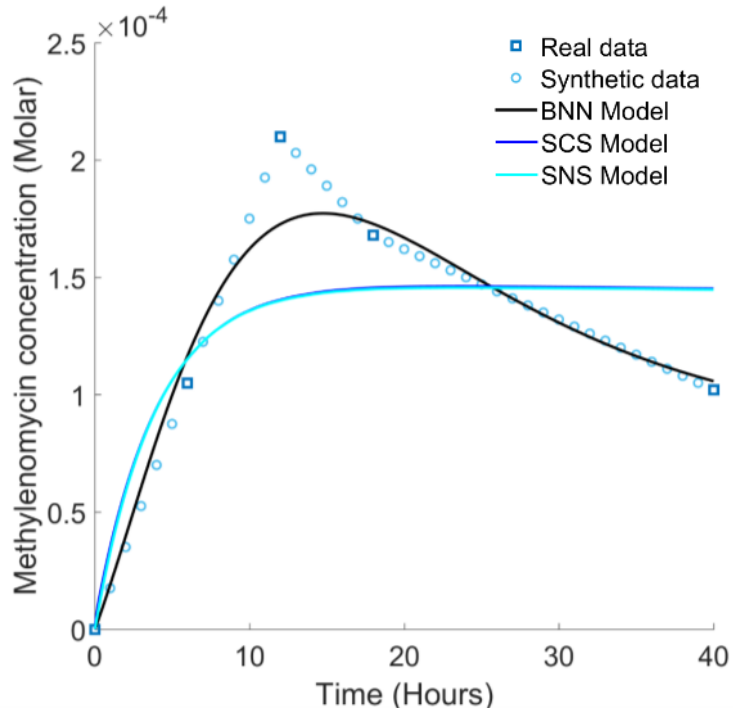


Figure 6.7: Genetic algorithm global optimisation results. The BNN model is able to fit the experimental data using the optimal parameter set identified by the GA. The absolute mean error provided by this optimal solution is  $6.12 \times 10^{-6}$ . The optimal fits provided by the SCS and SNS models are similar and are not as accurate as the BNN model. The absolute mean error provided by both of these optimal solutions is  $2.39 \times 10^{-5}$ .

data when optimised within the same parameter space used for model selection. The optimal parameter set identified by the GA is listed in Table 6.2 and provides an absolute mean error of  $6.12 \times 10^{-6}$ . The four parameter values describing protein

Reaction	Value ( $\text{M}^{-1}\text{s}^{-1}$ )	Reaction	Value ( $\text{s}^{-1}$ )
$k_3$	3.6119	$k_{-3}$	0.1092
$k_6$	0.9079	$k_{-6}$	5.7766
$k_{14}$	0.0065	$k_{-14}$	0.2208
—	—	$k_7$	2.6978
—	—	$k_8$	0.8902
—	—	$k_9$	5.8903
—	—	$k_{10}$	0.1101
—	—	$k_{11}$	0.6296
—	—	$k_{12}$	0.5307
—	—	$k_{13}$	0.0880
—	—	$\gamma_1$	0.9470
—	—	$\gamma_2$	2.7057
—	—	$\gamma_3$	0.2248
—	—	$\gamma_4$	0.1646

Table 6.2: Optimal parameter values for our BNN model. This optimal parameter set is dimensional, with parameters in the first and second reaction columns taking the units  $\text{M}^{-1}\text{s}^{-1}$  and  $\text{s}^{-1}$  respectively based on standard mass action kinetics.

degradation ( $\gamma_{1,2,3,4}$ ) vary by one order of magnitude at most; the remaining parameter values all describe protein:protein association and dissociation and vary by three orders of magnitude at most. Hence, we conclude that the numerical ranges of these optimal parameter values are reasonable within this biological context.

To investigate further, we also optimised the parameters of the SCS and SNS models against the experimental data using the GA, in an identical manner to that previously carried out for the BNN model. This revealed that neither model was able to achieve the same quality of fit to the data as the BNN (minimum error of  $2.39 \times 10^{-5}$  for both SCS and SNS compared to  $6.12 \times 10^{-6}$  for BNN). In addition, neither the SCS or SNS models were able to even qualitatively replicate the non-monotonicity in the response that is clearly exhibited in the experimental data.

## 6.6 Monte Carlo simulations of methylenomycin production in mutant strains

We perform additional model validation by testing the BNN model against our qualitative data regarding methylenomycin production in mutant *S. coelicolor* strains. We employ Monte Carlo simulations to examine methylenomycin production under four distinct conditions corresponding to the mutant strains described in Table 6.1. By examining the dynamical response to specific gene knockouts against the wild-type strain, represented by the optimal BNN model output in Fig. 6.7, we are able to investigate the qualitative effect of adapting our BNN model to emulate these

mutant strains.

When simulating MMY production in the different mutant strains, we account for  $\Delta mmyR$  by simply setting the parameter describing MmyR production from the  $fpm$ ,  $k_7$ , to 0. However,  $\Delta mmfR$  strains are incapable of producing MmfR and therefore cannot be simulated in the initial repressed state comprised of the  $fpm:MmfR$  and  $apm:MmfR$  complexes. Hence, the parameter describing MmyR production from the  $fpm$ ,  $k_8$ , is set to 0 and the allocation of initial concentrations is adapted to exclude the  $fpm:MmfR$  and  $apm:MmfR$  complexes. The  $\Delta mmfLHP$  strain is simulated by setting the initial concentration of the  $fpm$  and its associated complexes to 0, since the entire DNA module has been knocked out. The addition of exogenous MMF involves allocating this variable an initial concentration of 0.55 to align with the initial concentrations allocated to the relevant variables, that is, no new model parameters are introduced to describe production of exogenous MMF. Mutant strains comprising combinations of gene knockouts are simulated by combining the appropriate adaptations.

Specifically, in order to simulate the  $\Delta mmyR$  strain we set  $k_7 = 0$ . To simulate the  $\Delta mmfLHP$  strain the initial concentration of 0.55 is imposed on the  $apm:MmfR$  complex only, all other initial concentrations are set equal to 0. To simulate the  $\Delta mmfLHP + \Delta mmyR + \Delta mmfR$  strain we set  $k_7 = k_8 = 0$  and all initial concentration are set equal to 0 with the exception of the  $apm$  which is set equal to 0.55. To simulate the  $\Delta mmfLHP + \text{exogenous MMF}$  strain initial concentrations of the  $apm$  and MMF are set equal to 0.55, and all other initial concentrations are set equal to 0.

Monte Carlo simulations assign random values in the interval  $[10^{-4}, 10^4]$  to all model parameters, excluding those that retain their fixed values assigned for previous model selection and parameter inference purposes, as we continue to examine dimensional dynamic responses. We run a total of  $10^4$  Monte Carlo simulations to allow for substantial sampling of the parameter space within a feasible time frame. Each simulation outputs MMY production for each of the four mutant strains and calculates the ratio of the mean value to that of the optimal wildtype simulation. We utilise these ratios to investigate the ability of our model to satisfy the following four criteria, which capture the experimentally observed responses of the mutant strains:

1.  $\frac{\Delta mmyR}{\text{wildtype}} > 1.1$ ,
2.  $\frac{\Delta mmfLHP}{\text{wildtype}} < 0.9$ ,
3.  $\frac{\Delta mmfLHP + \Delta mmyR + \Delta mmfR}{\text{wildtype}} > 1.1$ ,



$$4. \ 0.9 < \frac{\Delta mmfLHP+MMF}{wildtype} < 1.1,$$

where over-production translates to an increase in mean MMY production of  $>10\%$ , cessation translates to a decrease in MMY production of  $>10\%$  and comparable production translates to a maximum increase or decrease in MMY production of less than  $10\%$ . The results of our Monte Carlo simulations are shown in Fig 6.8. Parameter sets were identified that are capable of satisfying each of the four criteria,

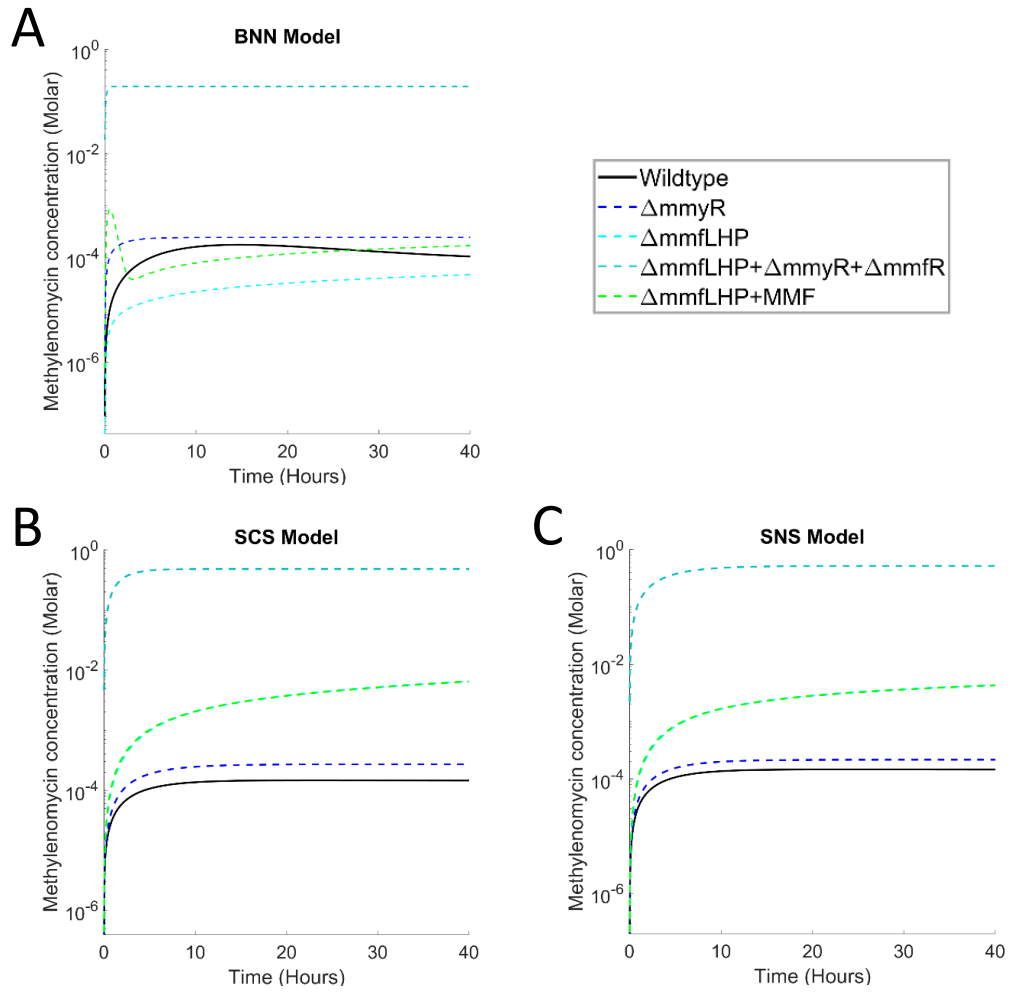


Figure 6.8: Monte Carlo simulation results. A) The BNN model is able to simulate the qualitative data regarding four mutant *S. coelicolor* strains when adapted to replicate the corresponding gene knockouts. B) The SCS model is unable to simulate qualitative data regarding the  $\Delta mmfLHP$  and  $\Delta mmfLHP+MMF$  knockout strains. C) The SNS model is unable to simulate qualitative data regarding the  $\Delta mmfLHP$  and  $\Delta mmfLHP+MMF$  knockout strains.

within the same dimensional solution space as the optimised wildtype model (Fig. 6.8A). Given the uncertainty regarding the effect of gene knockouts on the reaction kinetics and MMY production of the system, this qualitative agreement offers further validation of the replication and prediction capabilities of the BNN model.

The SCS and SNS models are also able to simulate the responses observed experimentally for the  $\Delta mmyR$  and the  $\Delta mmfLHP + \Delta mmyR + \Delta mmfR$  knockout strains, but not for the  $\Delta mmfLHP$  and  $\Delta mmfLHP + \text{exogenous MMF}$  knockouts (Fig. 6.8B and 6.8C). This is likely due to the most significant mechanistic property separating them from BNN, i.e. the interaction of MMY with one or both of MmyR and MmfR. This interaction results in decreased repression of the *apm* since the MMY is negating the action of either or both regulators and hence the *apm* is less restricted in producing MMY, which causes an over-production of the antibiotic for the  $\Delta mmfLHP$  and  $\Delta mmfLHP + \text{exogenous MMF}$  knockouts that has not been observed experimentally (Fig. 6.8B and 6.8C). We therefore conclude that the BNN model remains the most likely candidate model to explain all the available experimental data for this system.

## 6.7 Experimental design for future studies

We are able to inform the design of future experimental studies in light of our results. For example, we are interested in quantifying the response of the  $\Delta mmyR$  and  $\Delta mmfLHP + \Delta mmyR + \Delta mmfR$  strains in order to verify our model prediction that the five gene mutant elicits a more rapid and far greater over-production of MMY. This has implications both in terms of improving product yields for industrially relevant natural products, and also regarding the potential adverse effects this might cause in the cells such as toxicity. The result of these experiments would subsequently reveal whether the  $\Delta mmfLHP + \Delta mmyR + \Delta mmfR$  is the most effective knockout for improving antibiotic production in novel synthetic regulatory systems.

In the event that directly quantifying MMY production is inconclusive, we would be interested in replacing the gene controlled by the *apm* with a reporter gene coding for fluorescence or luminescence such as GFP or *lux* genes respectively. This output may enable us to measure the response of the different mutant strains with greater clarity since experiments of this nature are already well characterised, particularly in the related bacterium *S. venezuelae*.

Finally, we are also interested in examining the  $\Delta mmfLHP + \text{MMF}$  mutant in order to establish the quantity of exogenous MMF and the specific time point

of induction that provides optimal MMY production. Our model predicts a narrow production window for this strain which may suggest that direct MMY quantification is not straightforward and that, again, experimental designs incorporating reporter genes would provide improved results.

## 6.8 Conclusions

We have developed a plausible model architecture for the regulatory system controlling methylenomycin production in *Streptomyces coelicolor*. This architecture was found to be the most likely to reproduce the dynamical responses described by experimental time series data, when tested against 47 other candidate architectures of the same system. Global optimisation of the model parameters produced close agreement with the experimental data. Appropriate adjustments to the proposed model architecture allow it to replicate observed changes in the dynamics of methylenomycin production in a number of mutant *S. coelicolor* strains.

The mechanistic details captured in the proposed regulatory architecture provide useful insights for the design of future experiments to further investigate the operation of this system, and demonstrate the potential of mathematical models to elucidate the design principles of complex biological control systems. We expect that the emergence of further quantitative experimental data for this system will inform further model development and validation, and allow for the generation of optimised models that are capable of accurately predicting the dynamical responses of one of the most prevalent and important gene regulatory networks in nature.

## Chapter 7

# Conclusions and Future Work

### 7.1 Conclusions and discussion

The research presented in this thesis provides clear evidence of the importance of mathematical modelling approaches in synthetic biology. We have demonstrated that mechanistic models developed to reflect detailed biological observations can simulate and predict the dynamical responses of parts and devices, thus facilitating the engineering of novel biological systems.

Chapter 2 introduced the field of synthetic biology and outlined some of the techniques required to construct mechanistic mathematical models of biological systems. Mechanistic models are typically more complicated to derive and analyse in comparison with black box models, however this is often a small sacrifice to make for quantitative simulations and predictions. Construction of mathematical models is typically achieved through the application of mass action kinetics to systems of biochemical equations. The resulting system of ODEs can be solved explicitly using calculus although this is seldom applicable in practice due to the considerable dimensionality and non-linearity that often arises. Model ODEs are therefore commonly solved numerically *in silico* to provide approximate simulations of system dynamics. Model reduction is able to decrease dimensionality and non-linearity whilst preserving dynamical responses, however this is mainly beneficial in cases where rigorous mathematical analyses are prioritised. The dynamical response of biological systems is governed by the rates of the associated molecular interactions. Uncertainty regarding reaction rates is commonplace in synthetic biology, which often translates to large sets of unknown model parameter values. Two methods of parameter inference were demonstrated with respect to the same basic model, the first being the global optimisation technique known as the genetic algorithm and the second being

a Bayesian inference technique, ABC-SMC. Both methods were able to locate an optimal model parameter set in relation to a synthetic dataset. We demonstrated how a non-dimensional re-scaling of a mathematical model is another effective method of producing reliable simulations in light of parameter uncertainties.

Chapter 3 introduced DNA recombination and the significance of cellular memory in engineering novel synthetic systems. DNA recombination provides an ideal platform for cellular data storage and has formed the basis of the development of a rewritable RAD module, capable of efficient data storage within a chromosome. We have developed the first mechanistic model of DNA recombination, and validated it, through global optimisation, against a new set of *in vitro* data on recombination efficiencies across a range of different concentrations of integrase and gp3. We investigated *in vivo* recombination dynamics by exploiting non-dimensional model simulations. Our model revealed the importance of fully accounting for all mechanistic features of DNA recombination in order to accurately predict the effect of different switching strategies on RAD module performance, and highlighted the over-arching efficacy of models as design tools for building future synthetic circuitry.

The results of this modelling investigation confirm that DNA inversion events mediated solely by integrase proteins are highly efficient compared to dual-mediated inversion events that occur in the presence of integrase and gp3. Difficulties arise due to the residual concentration of integrase following an excision reaction which causes re-integration, rather than preserving the desired HOLD state. Our model also reveals two alternative strategies for overcoming such issues i.e. staggered cessation of integrase and gp3 induction, and constant integrase induction with switching efficiency as a function of gp3 induction. The latter strategy provides a step-like digital response, however neither of these solutions has practical viability since they both depend on extended induction of integrase or gp3, and the module is designed to hold state efficiently in the absence of induction. It is likely that new circuit designs will exploit inversion and deletion reactions mediated by integrase only until sufficient understanding of dual-mediated reactions is established.

The notion of a digital response may be misleading with regards to biological systems. It is unlikely that a step-like response can be induced biologically given that the time frames required for engineered cells to respond is relatively large, on the order of hours or even days. That said, our results serve as a promising proof of concept that suggests that the, more likely analogue, responses that arise from engineered genetic switches will be sufficient to elicit the same switching efficiency as their electronic counterparts on the required timescales.

Having demonstrated the efficacy of our mechanistic model against a sim-

plified alternative, it is clear that accounting for increased mechanistic detail is conducive to achieving reliable quantitative model outputs. That said, this simplified model is capable of providing qualitative simulations that are sufficient to direct experimental studies and its reduced dimensionality gives it the added advantage of presenting simplified mathematical analysis. Hence, although our modelling investigation does not prioritise rigorous mathematical analyses, model refinement is an important process that demands consideration in synthetic biology research.

Chapter 4 introduced Boolean logic functions and biological systems that exhibit analogous functionality. Recombinase-based circuitry that accounts for more than one protein input has been shown to enable the construction of circuits capable of temporal logic operations *in vivo*. Associated mathematical models had, to date, only captured the qualitative dynamical features of such systems and are thus of limited utility as tools to aid in the design of such circuitry. Our model of the RAD module was adapted in order to develop a detailed mechanistic model of a two-input temporal logic gate circuit based on unidirectional DNA recombination mediated by two distinct bacteriophage integrases, with the ability to detect and encode sequences of input events. We validated the model against *in vivo* experimental data through global optimisation and thus revealed quantitative replication and prediction of key dynamical features of the logic gate. Our model also predicts the effect of reversing integrase inputs on the output of the logic gate.

The results of this modelling investigation confirm that our model is capable of quantitative simulations of the temporal logic gate system, and therefore verify that the assembly of appropriate mechanisms comprising our validated model of the RAD module is an effective method for generating mechanistic models of higher-order circuitry. The preconception of the two integrase inputs, Bxb1 and TP901-1, was that their functional properties are very similar and could be assumed to be the same for modelling purposes. However, our model reveals that, in order for the logic gate circuit to produce the experimental data used to validate our model, there must be a notable difference in the rate of recombination mediated by the two integrases. Specifically, global optimisation identified the rate of recombination via Bxb1 to be  $\sim 1.8$ -fold greater than that of TP901-1. This has implications for the design of synthetic logic circuits since there is no guarantee that the expected functionality will be realised regardless of the specific roles assigned to each integrase input. Our model simulations confirmed that the reversal of integrase inputs has a significant impact on dynamical response, with both the steady state concentrations of fluorescent protein and the response time shown to increase as a result. Future modelling studies could potentially aim to characterise similar functional distinctions

between other integrases in order to provide synthetic biologists with the maximum amount of information regarding selection of integrase inputs for the functionality required. As the number of new, orthogonal integrases increases, a lookup table of responses might be useful to ensure that circuits function as expected.

Fitting our mechanistic model to the experimental data was successful in general, however datasets that captured a sigmoidal dynamical response presented the most difficulty with regard to matching model simulations. This highlights a limitation with the assumption that integrase expression can be represented mathematically by a constant parameter. Such constant parameters arise as a result of model derivation via mass action kinetics and are therefore valid mathematical terms, but this also restricts simulations to a particular form. The concentration of state variables will generally increase significantly within a short time frame due to constant inputs, making it difficult to replicate a more gradual, delayed increase without tuning specific parameter values. Hence, our model delivers greatest error for simulations of sigmoidal responses which suggests that greater consideration of the input expression parameters may be required to make improvements to these outputs. For example, time-dependent input expression parameters could be used to model additional dynamical intricacies, however the selected functions themselves, and any new associated parameters, would require conceptual validation with regards to biological systems. In all, we concluded that the error delivered by our model was minimised sufficiently to inform the design of recombinase-based logic circuitry without added complexities.

The notion of Boolean logic gate circuits may be misleading with respect to biological systems. Electronic logic circuits provide precise binary outputs dependent on precise binary inputs, and we do not claim that the biological logic circuits investigated in this thesis are capable of such performance. However, our results prove that recombinase-based logic circuits are capable of qualitative logic operations whereby induced inputs elicit transitioning to the associated outputs of the system in a similar manner to electronic circuits. Hence, logic gate is the term adopted throughout the literature to describe biologically equivalent circuits and is used throughout our research as a result.

Chapter 5 detailed the experimental project carried out to record *in vivo* DNA recombination efficiency. The integration reaction was not analysed due to its relative simplicity, however the efficiency of the excision reaction was quantified by recording the luminescence of the bacteria over time. Trial experiments were performed that identified the most suitable method of selecting bacteria that will grow efficiently and express their integrated genes in a predictable manner. Gen-

erally, however, bacterial growth was unpredictable and the resulting experimental data collated during the project were inconclusive. We concluded that validated mathematical models are invaluable tools in identifying and predicting dynamical behaviour, particularly in instances where the limitations of experimentation prohibit the collection of credible primary data.

A greatly important factor to consider with respect to experimental data and associated model simulations is the manner in which the data are processed prior to fitting. There is currently no standard practice for data processing given the considerable variety of experimental procedures that can be used to record biological system responses and mathematical modelling approaches that can be applied. In most cases, data must be normalised in order to fit a given model due to the fact that data collection is often performed using equipment that measures relative changes in outputs that are non-dimensional, and models derived on the basis of mass action kinetics retain dimensionality. Data collected *in vitro*, such as the data presented in Chapter 3, is potentially the most straightforward to process since the relevant biological quantities can be synthesised and used to normalise the resultant data. By contrast, data collected *in vivo*, such as the data presented in Chapter 4, can be more difficult to normalise since precise quantities of the reactants are often unobtainable. As synthetic biology research advances, there is likely to be increasing demand for standardised data processing techniques to broaden the acceptance of modelling investigations across the community. The development of procedures for the measurement of biological quantities *in vivo* will also be highly beneficial for establishing accurate conditions with which to initiate mathematical models.

Chapter 6 highlighted the historical significance of the discovery of antibiotics, and introduced the bacterium *Streptomyces coelicolor* as the model organism used to study the natural biosynthesis of the antibiotic methylenomycin A. Three candidate model architectures of the regulatory system controlling methylenomycin production were identified as the most plausible via an ABC-SMC model selection approach. The resultant mathematical models were optimised against an experimental time series dataset using the GA, revealing the strongest model architecture that can simulate the dynamical response of the system with minimal error. Monte Carlo simulations revealed distinct parameter sets capable of qualitatively replicating methylenomycin production in mutant *S. coelicolor* strains, based on a set of gene knockouts.

The results of this modelling investigation confirm that our current knowledge of the methylenomycin regulatory network is sufficient to inform the construction of a corresponding valid mathematical model. By identifying the most plausible



model architecture out of a set of 48 candidates, not only have we established a potential tool for the design of synthetic regulatory networks, but we have also revealed the likelihood of the existence of several structural mechanisms regarding the molecular interactions of certain regulatory elements. Further experimental studies will provide the opportunity for model development and verification of the plausible mechanistic properties that were statistically eliminated through model selection. The model was partially parameterised due to our kinetic data regarding the association and dissociation rates of the MmFR regulator to/from the MARE sequences. Performing specific experimental studies that run in tandem with modelling investigations by prioritising the measurement of kinetic data for model parameterisation was remarkably beneficial to our research efforts, and is likely to develop as a key goal for the field of synthetic biology.

In all, there can be no doubt that mathematical modelling approaches are deserving of a fundamental role in synthetic biology. Although we cannot claim that mathematical models are able to deliver perfect quantitative simulations of biological systems, the research presented in this thesis is evidence of the efficacy of models regarding the analysis of existing systems to assist the design of novel synthetic circuits. Synthetic biology remains a biology discipline at heart and, as such, modelling approaches are yet to receive the full attention and acceptance of the community as whole. Although there are currently multidisciplinary programs in place designed to produce new researchers with the full range of necessary skills, it may take several years before engineering and mathematical principles become totally integrated with experimental biology. In the meantime, it is likely that open-minded and proactive collaborations between biology and engineering labs will be essential in bridging existing interdisciplinary gaps, as well as realising a number of the primary goals of the field. Such collaborative efforts have contributed to the majority of the results presented in this thesis and will continue to strengthen the quality of the research carried out in subsequent studies.

## 7.2 Future work

The mechanistic nature of the models presented in this thesis has been shown to provide increased efficiency in comparison with simplistic models of the same systems. That said, the resultant complexity and dimensionality of our models has restricted our ability to provide full mathematical analyses. The application of model reduction was considered for our modelling investigations, particularly with respect to our model of the methylenomycin A producing gene cluster due to its

relatively small number of ODEs and parameters. However, reducing the model by two ODEs did not provide a sufficient reduction in mathematical investment to facilitate model analysis (not shown). We plan to undertake further work in identifying additional model reduction techniques to enable extensive mathematical analysis of our mechanistic models. A recent study by Hancock et al. [2015] presents a model reduction methodology that can be applied to gene regulation systems with scalability, and provides rigorous mathematical conditions on the relevant kinetic parameters. Reduction of a detailed gene regulation model preserves mechanistic output for just two dependent variables, the total concentration of mRNA and the total concentration of protein. When applied to the repressilator and toggle switch, the reduced models replicate the expected quantitative dynamical responses. This approach is developed in conjunction with transcriptional models of gene regulation, thus presenting very different dynamical mechanisms compared to our DNA recombination models. The proposed methodology is still applicable however, and may enable us, in future studies, to derive the mathematical criteria required of the desired functionality that we have identified.

Since the publication of the research presented in Chapter 3, a new model of DNA recombination reactions has been published [Pokhilko et al., 2016]. The model shares many similarities with our mechanistic model, being a deterministic ODE model constructed in light of structural analyses of serine integrases [Rutherford et al., 2013; van Duyne and Rutherford, 2013] that were included in our own literature mining. However, some significant differences are also reported. The new model is smaller dimensionally since less detail regarding DNA:protein binding is accounted for, for example the binding of monomeric protein is ignored. That said, greater detail regarding changes to the structural composition of the synaptic complexes as recombination occurs is modelled and therefore presents additional insight that could improve the output of our mechanistic model. The new model accounts for the formation of integrase:RDF tetramers in solution that are able to bind to free DNA however, we decided to exclude this mechanism from our model in light of literature mining and the depreciation in performance exhibited by models accounting for the mechanism. The new model also assumes that all system interactions are reversible, whereas we model the widely supported view that the recombination reactions themselves are unidirectional. We carried out an investigation into the reversibility of the excision reaction and observed a depreciation in model performance (see Chapter 3). The new model does not acknowledge the same formation of dysfunctional integrase dimers included in our mechanistic model. We are interested in examining the effect of these alternative mechanistic features on our model

outputs in order to improve the predictive capabilities of our design tools.

We also want to extend the investigation of temporal logic gate circuitry in Chapter 4 to infer the functional variations between distinct serine integrases. Our current results suggest that the two integrases used in implementing the temporal logic gate *in vivo* are not functionally identical since simultaneous induction of the integrases did not produce the expected 50% concentration of the final system state. Furthermore, we demonstrated that the reversal of integrase inputs caused considerable variation in the output of the logic gate (see Chapter 4). This suggests that our optimal parameter sets may infer new information regarding the rate of binding interactions for the two integrases, and also raises questions relating to output-specific induction schedules regarding a whole range of orthogonal integrases. That is, we would investigate whether our model could infer similar information for any pair of known integrases which might then enable us to establish a lookup table of dynamical outputs based on both the choice of the pair of integrase inputs, their associated induction schedule and their roles within the system.

Having validated mechanistic models of the recombinase-based genetic switch and temporal logic gate systems, the natural progression of this work is to consider the functional scope of more sophisticated logic gates and higher-level computing devices such as adders, subtractors and decoders. These devices implement multiple switches and logic gates and therefore present ideal platforms for extending our modelling investigations. We now plan to carry out an extensive modelling investigation regarding rewritable biocomputers in mammalian cells, together with our experimental collaborators at the Wong lab, Boston University. We will develop a 2-4 decoder, a foundational circuit topology that can form the basis of many other circuits. A 2-4 decoder takes two specific inputs and returns one of four distinct outputs. We have demonstrated how this can be achieved by the temporal logic gate (see Chapter 4) however, it can also be realised through standard logic functions with increased numbers of attachment site pairs. Our experimental collaborators have proposed two architectures for the decoder circuit design, one based on tyrosine recombination and one based on serine recombination. Both employ three pairs of specific attachment sites however, they employ just two distinct SSR inputs; one recombinase specific to two of the three pairs of sites and one recombinase specific to the remaining pair of sites. The tyrosine decoder is purely excision-based and is referred to as a Boolean Logic and Arithmetic through DNA Excision (BLADE) platform [Weinberg et al., 2017]. The serine decoder is purely inversion-based and exploits the recombinase binding interactions that have formed the focus of the research presented in Chapters 3 and 4. Constructing two models of the same system

that are distinguished by their recombinase mediation mechanisms will allow us to examine the comparative advantages and disadvantages concerning the choice of these systems to suit various applications.

The practical implementation of the biological decoder will require a detailed experimental analysis of serine integrase activity and specificity. Specifically, we will characterise Bxb1,  $\phi$ C31, TP901-1 and TG1 integrase systems in mammalian cell lines. We also plan to define the relationship between integrase expression level and specificity and toxicity as well as determine potential off-target sites that can cause harmful side effects. The design and characterisation of such inducible rewritable memory switches for the production of biopharmaceuticals will also require an exploration of the relationship between kinetic parameters and dose response profiles through integrated experimentation and mathematical modelling. To this end, we will characterise the performance of a range of memory switches, for example the T cell-stimulating factor interleukin-12 (IL-12) in CHO cells.

Since we already have validated serine integrase models that can potentially be extended or adapted to account for a wide variety of higher-level circuitry, the construction of a model of the serine decoder should be straightforward. However, we have yet to examine the mechanistic properties of tyrosine recombinase-based systems due to their comparatively limited functionality described in the literature. Future work will involve characterisation of genetic switches and logic gates that are mediated by tyrosine recombinases that we can then extend to tyrosine decoder models in the same manner that we have outlined for our mechanistic serine recombinase models.

# Bibliography

- C. Ajo-Franklin, D. Drubin, J. Eskin, E. Gee, D. Landgraf, I. Phillips, and P. Silver. Rational design of memory in eukaryotic cells. *Genes Dev*, 21(18):2271–76, 2007.
- B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular biology of the cell fourth edition*. Garland Science, 2002.
- U. Alon. *An introduction to systems biology: design principles of biological circuits*. Chapman and Hall/CRC, 2006.
- J. Anderson, E. Clarke, A. Arkin, and C. Voigt. Environmentally controlled invasion of cancer cells by engineered bacteria. *J Mol Biol*, 355(4):619–27, 2006.
- A. Arpino, E. Hancock, J. Anderson, M. Barahona, G-B. Stan, A. Papachristodoulou, and K. Polizzi. Tuning the dials of Synthetic Biology. *Microbiology*, 159(7):1236–53, 2013.
- M. Atkinson, M. Savageau, J. Myers, and A. Ninfa. Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in *Escherichia coli*. *Cell*, 113(5):597–607, 2003.
- S. Atsumi, T. Hanai, and J. Liao. Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. *Nature*, 451(7174):86–89, 2008.
- S. Auslander and M. Fussenegger. From gene switches to mammalian designer cells: present and future prospects. *Trends Biotechnol*, 31(3):155–68, 2013.
- H. Bai, M. Sun, P. Ghosh, G. Hatfull, N. Grindley, and J. Marko. Single-molecule analysis reveals the molecular bearing mechanism of DNA strand exchange by a serine recombinase. *Proc Natl Acad Sci USA*, 108(18):7419–24, 2011.
- R. Baltz. Strain improvement in actinomycetes in the postgenomic era. *J Ind Microbiol Biotechnol*, 38(6):657–66, 2011.

- R. Baltz. Streptomyces temperate bacteriophage integration systems for stable genetic engineering of actinomycetes (and other organisms). *J Ind Microbiol Biotechnol*, 39(5):661–72, 2012.
- S. Basu, Y. Gerchman, C. Collins, F. Arnold, and R. Weiss. A synthetic multicellular system for programmed pattern formation. *Nature*, 434(7037):1130–34, 2005.
- J. Baumgardner, K. Acker, O. Adefuye, S. Crowley, W. Deloache, J. Dickson, L. Heard, A. Martens, N. Morton, M. Ritter, A. Shoecraft, J. Treece, M. Unzicker, A. Valencia, M. Waters, A. Campbell, L. Heyer, J. Poet, and T. Eckdahl. Solving a Hamiltonian Path Problem with a bacterial computer. *J Biol Eng*, doi: 10.1186/1754-1611-3-11, 2009.
- T. Bayer and C. Smolke. Programmable ligand-controlled riboregulators of eukaryotic gene expression. *Nat Biotechnol*, 23(3):337–43, 2005.
- A. Becskei, B. Sraphin, and L. Serrano. Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. *EMBO J*, 20(10):2528–35, 2001.
- Y. Benenson, B. Gil, U. Ben-Dor, R. Adar, and E. Shapiro. An autonomous molecular computer for logical control of gene expression. *Nature*, 429(6990):423–29, 2004.
- S. Bentley, K. Chater, A. Cerdeño-Terraga, G. Challis, N. Thomson, K. James, D. Harris, M. Quail, H. Kieser, D. Harper, A. Bateman, S. Brown, G. Chandra, C. Chen, M. Collins, A. Cronin, A. Fraser, A. Goble, J. Hidalgo, T. Hornsby, S. Howarth, C. Huang, T. Kieser, L. Larke, L. Murphy, K. Oliver, S. O’Neil, E. Rabinowitsch, M. Rajandream, K. Rutherford, S. Rutter, K. Seeger, D. Saunders, S. Sharp, R. Squares, S. Squares, K. Taylor, T. Warren, A. Wietzorrek, J. Woodward, B. Barrell, J. Parkhill, and D. Hopwood. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature*, 417(6885):141–47, 2002.
- S. Bentley, S. Brown, L. Murphy, D. Harris, M. Quail, J. Parkhill, B. Barrell, J. McCormick, R. Santamaria, R. Losick, M. Yamasaki, H. Kinashi, C. Chen, G. Chandra, D. Jakimowicz, H. Kieser, T. Kieser, and K. Chater. SCP1, a 356 023 bp linear plasmid adapted to the ecology and developmental biology of its host, *Streptomyces coelicolor* A3(2). *Mol Microbiol*, 51(6):1615–28, 2004.
- L. Bibb and G. Hatfull. Integration and excision of the *Mycobacterium tuberculosis* prophage-like element,  $\phi$ Rv1. *Mol Microbiol*, 45(6):1515–26, 2002.

- L. Bibb, M. Hancox, and G. Hatfull. Integration and excision by the large serine recombinase phiRv1 integrase. *Mol Microbiol*, 55(6):1896–910, 2005.
- J. Bonnet, P. Subsoontorn, and D. Endy. Rewritable digital data storage in live cells via engineered control of recombination directionality. *Proc Natl Acad Sci USA*, 109(23):8884–89, 2012.
- J. Bonnet, P. Yin, M. Ortiz, P. Subsoontorn, and D. Endy. Amplifying Genetic Logic Gates. *Science*, 340(6132):599–603, 2013.
- J. Bowyer, J. Zhao, S. Rosser and S. Colloms, and D. Bates. Development and experimental validation of a mechanistic model of *in vitro* DNA recombination. In *Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 945–48, 2015.
- J. Bowyer, J. Zhao, P. Subsoontorn, W. Wong, S. Rosser, and D. Bates. Mechanistic Modeling of a Rewritable Recombinase Addressable Data Module. *IEEE Transactions on Biomedical Circuits and Systems*, 10(6):1161–70, 2016.
- C. Branda and S. Dymecki. Talking about a revolution: The impact of site-specific recombinases on genetic analyses in mice. *Dev Cell*, 6(1):7–28, 2004.
- A. Breuner, L. Brondsted, and K. Hammer. Novel organization of genes involved in prophage excision identified in the temperate lactococcal bacteriophage TP901-1. *Bacteriol*, 181(23):7291–97, 1999.
- A. Breuner, L. Brondsted, and K. Hammer. Resolvase-like recombination performed by the TP901-1 integrase. *Microbiology*, 147(8):2051–63, 2001.
- K. Brown. *Penicillin Man: Alexander Fleming and the Antibiotic Revolution*. Sutton Publishing, 2005.
- W. Brown, N. Lee, Z. Xu, and M. Smith. Serine recombinases as tools for genome engineering. *Methods*, 53(4):372–79, 2011.
- F. Buchholz, L. Ringrose, P. Angrand, F. Rossi, and A. Stewart. Different thermostabilities of FLP and Cre recombinases: implications for applied site-specific recombination. *Nucleic Acids Res*, 24(21):4256–62, 1996.
- D. Burrill and P. Silver. Making cellular memories. *Cell*, 140(1):13–18, 2010.
- J. Callura, D. Dwyer, F. Isaacs, C. Cantor, and J. Collins. Tracking, tuning, and terminating microbial physiology using synthetic riboregulators. *Proc Natl Acad Sci USA*, 107(36):15898–903, 2010.

- D. Cameron, C. Bashor, and J. Collins. A brief history of synthetic biology. *Nat Rev Microbiol*, 12(5):381–90, 2014.
- K. Chater and C. Bruton. Resistance, regulatory and production genes for the antibiotic methylenomycin are clustered. *EMBO J*, 4(7):1893–97, 1985.
- C. Chavez, A. Keravala, L. Woodard, R. Hillman, T. Stowe, J. Chu, and M. Calos. Kinetics and longevity of C31 integrase in mouse liver and cultured cells. *Hum Gene Ther*, 21(10):1287–97, 2010.
- B-S. Chen and P-W. Chen. GA-based design algorithms for the robust synthetic genetic oscillators with prescribed amplitude, period and phase. *Gene Regul Syst Bio*, 4:35–52, 2010.
- W. Chen and Z. Qin. Development of a gene cloning system in a fast-growing and moderately thermophilic *Streptomyces* species and heterologous expression of *Streptomyces* antibiotic biosynthetic gene clusters. *BMC Microbiol*, doi:10.1186/1471-2180-11-243, 2011.
- E. Cho, R. Gumport, and J. Gardner. Interactions between integrase and excisionase in the phage lambda excisive nucleoprotein complex. *J Bacteriol*, 184(18):5200–03, 2002.
- G. Church, M. Elowitz, C. Smolke, C. Voigt, and R. Weiss. Realizing the potential of synthetic biology. *Nat Rev Mol Cell Biol*, 15(4):289–94, 2014.
- S. Colloms, C. Merrick, F. Olorunniji, M. Stark, M. Smith, A. Osbourn, J. Keasling, and S. Rosser. Rapid metabolic pathway assembly and modification using serine integrase site-specific recombination. *Nucleic Acids Res*, 42(4):e23. doi:10.1093/nar/gkt1101, 2014.
- P. Combes, R. Till, S. Bee, and M. Smith. The streptomyces genome contains multiple pseudo-attB sites for the (phi)C31-encoded site-specific recombination system. *J Bacteriol*, 184(20):5746–52, 2002.
- C. Corre. In search of the missing ligands for TetR family regulators. *Chem Biol*, 20(2):140–42, 2013.
- C. Corre and G. Challis. Evidence for the unusual condensation of a diketide with a pentose sugar in the methylenomycin biosynthetic pathway of *Streptomyces coelicolor* A3(2). *ChemBioChem*, 6:2166–70, 2005.



- C. Corre, L. Song, S. O'Rourke, K. Chater, and G. Challis. 2-Alkyl-4-hydroxymethylfuran-3-carboxylic acids, antibiotic production inducers discovered by *Streptomyces coelicolor* genome mining. *Proc Natl Acad Sci USA*, 105(45):17510–15, 2008.
- J. Davies and D. Davies. Origins and Evolution of Antibiotic Resistance. *Microbiol Mol Biol Rev*, 74(3):417–33, 2010.
- M. DuPage, A. Dooley, and T. Jacks. Conditional mouse lung cancer models using adenoviral or lentiviral delivery of Cre recombinase. *Nat Protoc*, 4(7):1064–72, 2009.
- M. Elowitz and S. Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–38, 2000.
- M. Fernandez, J. Caballero, L. Fernandez, and A. Sarai. Genetic algorithm optimization in drug design QSAR: Bayesian-regularized genetic neural networks (BRGNN) and genetic algorithm-optimized support vectors machines (GA-SVM). *Mol Divers*, 15(1):269–89, 2011.
- J. Fernandez-Rodriguez, L. Yang, T. Gorochoewski, D. Gordon, and C. Voigt. Memory and combinatorial logic based on DNA inversions: dynamics and evolutionary stability. *ACS Synth Biol*, 4(12):1361–72, 2015.
- M. Finland. Emergence of antibiotic resistance in hospitals, 1935-1975. *Rev Infect Dis*, 1(1):4–22, 1979.
- K. Flårdh and M. Buttner. *Streptomyces* morphogenetics: dissecting differentiation in a filamentous bacterium. *Nat Rev Microbiol*, 7(1):36–49, 2009.
- A. Fleming. Classics in infectious diseases: on the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of *B. influenzae* by Alexander Fleming, Reprinted from the British Journal of Experimental Pathology 10:226-236, 1929. *Rev Infect Dis*, 2(1):129–39, 1980.
- P. Fogg, S. Colloms, S. Rosser, M. Stark, and M. Smith. New applications for phage integrases. *J Mol Biol*, 426(15):2703–16, 2014.
- A. Friedland, T. Lu, X. Wang, D. Shi, G. Church, and J. Collins. Synthetic gene networks that count. *Science*, 324(5931):1199–202, 2009.
- T. Gardner, C. Cantor, and J. Collins. Construction of a genetic toggle switch in *Escherichia coli*. *Nature*, 403(6767):339–42, 2000.

- N. Gerber and H. Lechevalier. Geosmin, an earthy-smelling substance isolated from actinomycetes. *Appl Microbiol*, 13(6):935–38, 1965.
- P. Ghosh, A. Kim, and G. Hatfull. The orientation of mycobacteriophage Bxb1 integration is solely dependent on the central dinucleotide of attP and attB. *Mol Cell*, 12(5):1101–11, 2003.
- P. Ghosh, N. Pannunzio, and G Hatfull. Synapsis in phage Bxb1 integration: selection mechanism for the correct pair of recombination sites. *J Mol Biol*, 349(2):331–48, 2005.
- P. Ghosh, L. Wasil, and G. Hatfull. Control of phage Bxb1 excision by a novel recombination directionality factor. *PLoS Biol*, 4(6):e186, 2006.
- P. Ghosh, L. Bibb, and G. Hatfull. Two-step site selection for serine-integrase-mediated excision: DNA-directed integrase conformation and central dinucleotide proofreading. *Proc Natl Acad Sci USA*, 105(9):3238–43, 2008.
- D. Gillies. Lecture 1: an introduction to Boolean algebra. <http://www.doc.ic.ac.uk/~dfg/hardware/HardwareLecture01.pdf>, 2010.
- K. Goh, B. Kahng, and K. Cho. Sustained oscillations in extended genetic oscillatory systems. *Biophys J*, 94(11):4270–76, 2008.
- M. Gregory, R. Till, and M. Smith. Integration site for Streptomyces phage phiBT1 and development of site-specific integrating vectors. *J Bacteriol*, 185(17):5320–23, 2003.
- N. Grindley, K. Whiteson, and P. Rice. Mechanisms of site-specific recombination. *Annu Rev Biochem*, 75:567–605, 2006.
- A. Groth and M. Calos. Phage integrases: biology and applications. *J Mol Biol*, 335(3):667–78, 2004.
- M. Guinn and L. Bleris. Biological 2-input decoder circuit in human cells. *ACS Synth Biol*, 3(8):627–33, 2014.
- M. Gupta, R. Till, and M. Smith. Sequences in attB that affect the ability of phiC31 integrase to synapse and to activate DNA cleavage. *Nucleic Acids Res*, 35(10):3407–19, 2007.
- A. Haldimann and B. Wanner. Conditional-replication, integration, excision, and retrieval plasmid-host systems for gene structure-function studies of bacteria. *J. Bacteriol*, 183(21):6384–6393, 2001.

- T. Ham, S. Lee, J. Keasling, and A. Arkin. Design and construction of a double inversion recombination switch for heritable sequential genetic memory. *PLoS One*, 3(7) e2815. doi: 10.1371/journal.pone.0002815, 2008.
- E. Hancock, G-B. Stan, J. Arpino, and A. Papachristodoulou. Simplified mechanistic models of gene regulation for analysis and design. *J R Soc Interface*, 12(108), 2015.
- R. Harrison. Introduction to monte carlo simulation. *AIP Conf Proc*, 1204:17–21, 2010.
- A. Hayes, G. Hobbs, C. Smith, S. Oliver, and P. Butler. Environmental signals triggering methylenomycin production by *Streptomyces coelicolor* A3(2). *J Bacteriol*, 179(17):5511–15, 1997.
- E. Hendrix and B. Gazdag-Toth. *Introduction to nonlinear and global optimization*. Springer, 2010.
- G. Hobbs, A.Obanye, J. Petty, J. Mason, E. Barratt, D. Gardner, F. Flett, C. Smith, P. Broda, and S. Oliver. An integrated approach to studying regulation of production of the antibiotic methylenomycin by *Streptomyces coelicolor* A3(2). *J Bacteriol*, 174(5):1487–94, 1992.
- V. Hsiao, E. de los Santos, W. Whitaker, J. Dueber, and R. Murray. Design and implementation of a biomolecular concentration tracker. *ACS Synth Biol*, 4(2): 150–61, 2015.
- V. Hsiao, Y. Hori, P. Rothmund, and R. Murray. A population-based temporal logic gate for timing and recording chemical events. *Mol Syst Biol*, 12(5):869, 2016.
- G. Hu, M. Goll, and S. Fisher.  $\phi$ C31 integrase mediates efficient cassette exchange in the zebrafish germline. *Dev Dyn*, 240(9):2101–07, 2011.
- F. Isaacs, J. Hasty, C. Cantor, and J. Collins. Prediction and measurement of an autoregulatory genetic module. *Proc Natl Acad Sci USA*, 100(13):7714–19, 2003.
- D. Jakimowicz and G van Wezel. Cell division and DNA segregation in *Streptomyces*: how to build a septum in the middle of nowhere? *Mol Microbiol*, 85(3): 393–404, 2012.
- E. Kapusi, K. Kempe, M. Rubtsova, J. Kumlehn, and M. Gils.  $\phi$ C31 integrase-mediated site-specific recombination in barley. *PLoS One*, 7(9):e45353, 2012.

- R. Keenholtz, S. Rowland, M. Boocock, M. Stark, and P. Rice. Structural basis for catalytic activation of a serine recombinase. *Structure*, 19(6):799–809, 2011.
- K. Kempe, M. Rubtsova, C. Berger, J. Kumlehn, C. Schollmeier, and M. Gils. Transgene excision from wheat chromosomes by phage phiC31 integrase. *Plant Mol Biol*, 72(6):673–87, 2010.
- A. Keravala, J. Portlock, J. Nash, D. Vitrant, P. Robbins, and M. Calos. PhiC31 integrase mediates integration in cultured synovial cells and enhances gene expression in rabbit joints. *J Gene Med*, 8(8):1008–17, 2006.
- M. Kershaw, J. Westwood, and P. Darcy. Gene-engineered T cells for cancer therapy. *Nat Rev Cancer*, 13(8):525–41, 2013.
- T. Khaleel, E. Younger, A. McEwan, A. Varghese, and M. Smith. A phage protein that binds  $\phi$ C31 integrase to switch its directionality. *Molecular Microbiology*, 80(6):1450–63, 2011.
- T. Kieser, M. Bibb, M. Buttner, K. Chater, and D. Hopwood. *Practical Streptomyces Genetics*. John Innes Foundation, 2000.
- N. Kilby, M. Snaith, and J. Murray. Site-specific recombinases: tools for genome engineering. *Trends Genet*, 9(12):413–21, 1993.
- A. Kim, P. Ghosh, M. Aaron, L. Bibb, S. Jain, and G. Hatfull. Mycobacteriophage Bxb1 integrates into the Mycobacterium smegmatis groEL1 gene. *Mol Microbiol*, 50(2):463–73, 2003.
- E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach. *Systems biology in practice: concepts, implementation and application*. Wiley-VCH, 2005.
- H. Kobayashi, M. Kaern, M. Araki, K. Chung, T. Gardner, C. Cantor, and J. Collins. Programmable cells: interfacing natural and engineered gene networks. *Proc Natl Acad Sci USA*, 101(22):8414–19, 2004.
- B. Kramer, W. Weber, and M. Fussenegger. Artificial regulatory networks and cascades for discrete multilevel transgene control in mammalian cells. *Biotechnol Bioeng*, 83(7):810–20, 2003.
- B. Kramer, C. Fischer, and M. Fussenegger. BioLogic gates enable logical transcription control in mammalian cells. *Biotechnol Bioeng*, 87(4):478–84, 2004a.

- B. Kramer, A. Viretta, M. Daoud-El-Baba, D. Aubel, W. Weber, and M. Fussenegger. An engineered epigenetic transgene switch in mammalian cells. *Nat Biotechnol*, 22(7):867–70, 2004b.
- A. Landy. The  $\lambda$  integrase site-specific recombination pathway. *Microbiol Spectr*, 3(2). doi: 10.1128/microbiolspec.MDNA3-0051-2014, 2015.
- L. Laureti, L. Song, S. Huang, C. Corre, P. Leblond, G. Challis, and B. Aigle. Identification of a bioactive 51-membered macrolide complex by activation of a silent polyketide synthase in *Streptomyces ambofaciens*. *Proc Natl Acad Sci USA*, 108(15):6258–63, 2011.
- J. Lewis and G. Hatfull. Identification and characterization of mycobacteriophage L5 excisionase. *Mol Microbiol*, 35(2):350–60, 2000.
- F. Lienert, J. Lohmueller, A. Garg, and P. Silver. Synthetic biology in mammalian cells: next generation research tools and therapeutics. *Nat Rev Mol Cell Biol*, 15(2):95–107, 2014.
- J. Liepe, P. Kirk, S. Filippi, T. Toni, C. Barnes, and M. Stumpf. A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. *Nat Protoc*, 9(2):439–56, 2014.
- J. Lister. Transgene excision in zebrafish using the phiC31 integrase. *Genesis*, 48(2):137–43, 2010.
- G. Liu, K. Chater, G. Chandra, G. Niu, and H. Tan. Molecular regulation of antibiotic biosynthesis in streptomyces. *Microbiol Mol Biol Rev*, 77(1):112–43, 2013.
- S. Liu, J. Ma, W. Wang, M. Zhang, Q. Xin, S. Peng, R. Li, and H. Zhu. Mutational analysis of highly conserved residues in the phage phiC31 integrase reveals key amino acids necessary for the DNA recombination. *PLoS One*, 5(1) e8863. doi: 10.1371/journal.pone.0008863, 2010.
- J. Lohmueller, T. Armel, and P. Silver. A tunable zinc finger-based framework for Boolean logic computation in mammalian cells. *Nucleic Acids Res*, 40(11):5180–87, 2012.
- C. Lou, X. Liu, M. Ni, Y. Huang, Q. Huang, L. Huang, L. Jiang, D. Lu, M. Wang, C. Liu, D. Chen, C. Chen, X. Chen, L. Yang, H. Ma, J. Chen, and Q. Ouyang. Synthesizing a novel genetic sequential logic circuit: a push-on push-off switch. *Mol Syst Biol*, doi: 10.1038/msb.2010.2, 2010.

- T. Lu and J. Collins. Dispersing biofilms with engineered enzymatic bacteriophage. *Proc Natl Acad Sci USA*, 104(27):11197–202, 2007.
- I. Lucet, F. Tynan, V. Adams, J. Rossjohn, D. Lyras, and J. Rood. Identification of the structural and functional domains of the large serine recombinase TnpX from *Clostridium perfringens*. *J Biol Chem*, 280(4):2503–11, 2005.
- S. Mandali, G. Dhar, N. Avliyakov, M. Haykinson, and R. Johnson. The site-specific integration reaction of *Listeria* phage A118 integrase, a serine recombinase. *Mob DNA*, 4(1). doi: 10.1186/1759-8753-4-2, 2013.
- W. Marshall, K. Young, M. Swaffer, E. Wood, P. Nurse, A. Kimura, J. Frankel, J. Wallingford, V. Walbot, X. Qu, and A. Roeder. What determines cell size? *BMC Biol*, 10(101). doi: 10.1186/1741-7007-10-101, 2012.
- V. Martin, D. Pitera, S. Withers, J. Newman, and J. Keasling. Engineering a mevalonate pathway in *Escherichia coli* for production of terpenoids. *Nat Biotechnol*, 21(7):796–802, 2003.
- M. Matsuura, T. Noguchi, D. Yamaguchi, T. Aida, M. Asayama, H. Takahashi, and M. Shirai. The sre gene (ORF469) encodes a site-specific recombinase responsible for integration of the R4 phage genome. *J Bacteriol*, 178(11):3374–76, 1996.
- A. McEwan, P. Rowley, and M. Smith. DNA binding and synapsis by the large C-terminal domain of phiC31 integrase. *Nucleic Acids Res*, 37(14):4764–73, 2009.
- A. McEwan, A. Raab, S. Kelly, J. Feldmann, and M. Smith. Zinc is essential for high-affinity DNA binding and recombinase activity of  $\phi$ C31 integrase. *Nucleic Acids Res*, 39(14):6137–47, 2011.
- J-B. Michel, P. Yeh, R. Chait, R. Moellering, and R. Kishony. Drug interactions modulate the potential for evolution of resistance. *Proc Natl Acad Sci USA*, 105(39):14918–23, 2008.
- T. Miura, Y. Hosaka, Y. Yan-Zhuo, T. Nishizawa, M. Asayama, H. Takahashi, and M. Shirai. *In vivo* and *in vitro* characterization of site-specific recombination of actinophage R4 integrase. *J Gen Appl Microbiol*, 57(1):45–57, 2011.
- G. Moe-Behrens. The biological microprocessor, or how to build a computer with biological parts. *Comput Struct Biotechnol J*, 7 e201304003. doi: 10.5936/csbj.201304003, 2013.

- T. Moon, C. Lou, B. Stanton A. Tamsir, and C. Voigt. Genetic programs constructed from layered logic gates in single cells. *Nature*, 491(7423):249–53, 2012.
- M. Morange. The scientific legacy of Jacques Monod. *Res Microbiol*, 161(2):77–81, 2010.
- M. Morange. Francois Jacob (1920-2013). *Nature*, 497(7450):440, 2013.
- J. Murray. *Mathematical biology I: an introduction*. Springer, 2002.
- D. Nicholl. *An introduction to genetic engineering*. Cambridge University Press, 2008.
- L. Nkrumah, R. Muhle, P. Moura, P. Ghosh, G. Hatfull, W. Jacobs, and D. Fidock. Efficient site-specific integration in *Plasmodium falciparum* chromosomes mediated by mycobacteriophage Bxb1 integrase. *Nat Methods*, 3(8):615–21, 2006.
- A. O’Driscoll and R. Sleator. Synthetic DNA: the next generation of big data storage. *Bioengineered*, 4(3):123–5, 2013.
- E. Olivares, R. Hollis, and M. Calos. Phage R4 integrase mediates site-specific integration in human cells. *Gene*, 278(1-2):167–76, 2001.
- F. Olorunniji and M. Stark. Catalysis of site-specific recombination by Tn3 resolvase. *Biochem Soc Trans*, 38(2):417–21, 2010.
- F. Olorunniji, D. Buck, S. Colloms, A. McEwan, M. Smith, M. Stark, and S. Rosser. Gated rotation mechanism of site-specific recombination by  $\phi$ C31 integrase. *Proc Natl Acad Sci USA*, 109(48):19661–6, 2012.
- S. O’Rourke, A. Wietzorrek, K. Fowler, C. Corre, G. Challis, and K. Chater. Extracellular signalling, translational control, two repressors and an activator all contribute to the regulation of methylenomycin production in *Streptomyces coelicolor*. *Mol Microbiol*, 71(3):763–78, 2009.
- S. Ortiz-Urda, B. Thyagarajan, D. Keene, Q. Lin, M. Fang, M. Calos, and P. Khavari. Stable nonviral genetic correction of inherited human skin disease. *Nat Med*, 8(10):1166–70, 2002.
- S. Ortiz-Urda, B. Thyagarajan, D. Keene, Q. Lin, M. Calos, and P. Khavari. PhiC31 integrase-mediated nonviral genetic correction of junctional epidermolysis bullosa. *Hum Gene Ther*, 14(9):923–28, 2003.

- C. Paddon, P. Westfall, D. Pitera, K. Benjamin, K. Fisher, D. McPhee, M. Leavell, A. Tai, A. Main, D. Eng, D. Polichuk, K. Teoh, D. Reed, T. Treynor, J. Lenihan, M. Fleck, S. Bajad, G. Dang, D. Dengrove, D. Diola, G. Dorin, K. Ellens, S. Fickes, J. Galazzo, S. Gaucher, T. Geistlinger, R. Henry, M. Hepp, T. Horning, T. Iqbal, H. Jiang, L. Kizer, B. Lieu, D. Melis, N. Moss, R. Regentin, S. Secrest, H. Tsuruta, R. Vazquez, L. Westblade, L. Xu, M. Yu, Y. Zhang, L. Zhao, J. Lievense, P. Covello, J. Keasling, K. Reiling, N. Renninger, and J. Newman. High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature*, 496(7446):528–32, 2013.
- S. Park, A. Zarrinpar, and W. Lim. Rewiring MAP kinase pathways using alternative scaffold assembly mechanisms. *Science*, 299(5609):1061–64, 2003.
- C. Pena, J. Kahlenberg, and G. Hatfull. Protein-DNA complexes in mycobacteriophage L5 integrative recombination. *J Bacteriol*, 181(2):454–61, 1999.
- A. Pokhilko, J. Zhao, O. Ebenhöf, M. Smith, M. Stark, and S. Colloms. The mechanism of C31 integrase directionality: experimental analysis and computational modelling. *Nucleic Acids Res*, 44(15):7360–72, 2016.
- M. Ptashne. *A genetic switch*. Cold Spring Harbor Press; 3rd revised edition, 2004.
- M. Ptashne. Lambda’s switch: lessons from a module swap. *Curr Biol*, 16(12):459–62, 2006.
- O. Purcell and T. Lu. Synthetic analog and digital circuits for cellular computation and memory. *Curr Opin Biotechnol*, doi: 10.1016/j.copbio.2014.04.009, 2014.
- P. Purnick and R. Weiss. The second wave of synthetic biology. *Nat Rev Mol Cell Biol*, 10(6):410–22, 2009.
- O. Rackham and J. Chin. Cellular logic with orthogonal ribosomes. *J Am Chem Soc*, 127(50):17584–85, 2005.
- J. Ramos, M. Martnez-Bueno, A. Molina-Henares, W. Tern, K. Watanabe, X. Zhang, M. Gallegos, R. Brennan, and R. Tobes. The TetR family of transcriptional repressors. *Microbiol Mol Biol Rev*, 69(2):326–56, 2005.
- M. Rashel, J. Uchiyama, T. Ujihara, I. Takemura, H. Hoshiba, and S. Matsuzaki. A novel site-specific recombination system derived from bacteriophage phiMR11. *Biochem Biophys Res Commun*, 368(2):192–98, 2008.



- S. Regot, J. Macia, N. Conde, K. Furukawa, J. Kjelln, T. Peeters, S. Hohmann, E de Nadal, F. Posas, and R. Sol. Distributed biological computation with multicellular engineered networks. *Nature*, 469(7329):207–11, 2011.
- K. Rinaudo, L. Bleris, R. Maddamsetti, S. Subramanian, R. Weiss, and Y. Benenson. A universal RNAi-based logic evaluator that operates in mammalian cells. *Nat Biotechnol*, 25(7):795–801, 2007.
- L. Ringrose, V. Lounnas, L. Ehrlich, F. Buchholz, R. Wade, and A. Stewart. Comparative kinetic analysis of FLP and cre recombinases: mathematical models for DNA binding and recombination. *J Mol Biol*, 284(2):363–84, 1998.
- D. Ro, E. Paradise, M. Ouellet, K. Fisher, K. Newman, J. Ndungu, K. Ho, R. Eachus, T. Ham, J Kirby, M. Chang, S. Withers, Y. Shiba, R. Sarpong, and J. Keasling. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, 440(7086):940–43, 2006.
- S. Ross. *A first course in probability eighth edition*. Pearson Prentice Hall, 2010.
- J. Rossant and A. Nagy. Genome engineering: the new mouse genetics. *Nat Med*, 1(6):592–4, 1995.
- P. Rowley and M. Smith. Role of the N-terminal domain of phiC31 integrase in attB-attP synapsis. *J Bacteriol*, 190(20):6918–21, 2008.
- K. Rutherford, P. Yuan, K. Perry, R. Sharp, and G van Duyne. Attachment site recognition and regulation of directionality by the serine integrases. *Nucleic Acids Res*, 41(17):8341–56, 2013.
- D. Shis, F. Hussain, S. Meinhardt, L. Swint-Kruse, and M. Bennett. Modular, multi-input transcriptional logic gating with orthogonal LacI/GalR family chimeras. *ACS Synth Biol*, 3(9):645–51, 2014.
- J. Sidda, L. Song, V. Poon, M. Al-Bassam, O. Lazos, M. Buttner, G. Challis, and C. Corre. Discovery of a family of  $\gamma$ -aminobutyrate ureas via rational derepression of a silent bacterial gene cluster. *Chem Sci*, S(1):86–89, 2014.
- S. Singh, P. Ghosh, and G. Hatfull. Attachment site selection and identity in Bxb1 serine integrase-mediated site-specific recombination. *PLoS Genet*, 9(5) e1003490. doi: 10.1371/journal.pgen.1003490, 2013.
- P. Siuti, J. Yazbek, and T. Lu. Synthetic circuits integrating logic and memory in living cells. *Nat Biotechnol*, 31(5):448–52, 2013.

- M. Smith and H. Thorpe. Diversity in the serine recombinases. *Mol Microbiol*, 44(2):299–307, 2002.
- M. Smith, R. Till, K. Brady, P. Soultanas, H. Thorpe, and M. Smith. Synapsis and DNA cleavage in phiC31 integrase-mediated site-specific recombination. *Nucleic Acids Res*, 32(8):2607–17, 2004.
- M. Smith, W. Brown, A. McEwan, and P. Rowley. Site-specific recombination by phiC31 integrase and other large serine recombinases. *Biochem Soc Trans*, 38(2):388–94, 2010.
- R. Smith and Y. Grohn. Use of approximate Bayesian computation to assess and fit models of mycobacterium leprae to predict outcomes of the Brazilian control program. *PLoS One*, 10(6) e0129535. doi: 10.1371/journal.pone.0129535, 2015.
- B. Stanton, A. Nielsen, A. Tamsir, K. Clancy, T. Peterson, and C. Voigt. Genomic mining of prokaryotic repressors for orthogonal logic gates. *Nat Chem Biol*, 10(2):99–105, 2014.
- M. Stark, M. Boocock, F. Olorunniji, and S. Rowland. Intermediates in serine recombinase-mediated site-specific recombination. *Biochem Soc Trans*, 39(2):617–22, 2011.
- M. Stojanovic and D. Stefanovic. A deoxyribozyme-based molecular automaton. *Nat Biotechnol*, 21(9):1069–74, 2003.
- J. Stricker, S. Cookson, M. Bennett, W. Mather, L. Tsimring, and J. Hasty. A fast, robust and tunable synthetic gene oscillator. *Nature*, 456(7221):516–9, 2008.
- M. Stumpf. Approximate Bayesian inference for complex ecosystems. *F1000Prime Rep*, doi: 10.12703/P6-60, 2014.
- K. Styles. *Investigating interactions between methylenomycin furan microbial hormones and transcriptional repressors in Streptomyces coelicolor*. PhD thesis, School of Life Sciences, University of Warwick, 2016.
- B. Swalla, E. Cho, R. Gumport, and J. Gardner. The molecular basis of co-operative DNA binding between lambda integrase and excisionase. *Mol Microbiol*, 50(1):89–99, 2003.
- I. Swinburne, D. Miguez, D. Landgraf, and P. Silver. Intron length increases oscillatory periods of gene expression in animal cells. *Genes Dev*, 22(17):2342–46, 2008.

- J. Tabor, H. Salis, Z. Simpson, A. Chevalier, A. Levskaya, E. Marcotte, C. Voigt, and A. Ellington. A synthetic genetic edge detection program. *Cell*, 137(7):1272–81, 2009.
- A. Tamsir, J. Tabor, and C. Voigt. Robust multicellular computing using genetically encoded NOR gates and chemical ‘wires’. *Nature*, 469(7329):212–15, 2011.
- B. Tasic, S. Hippenmeyer, C. Wang, M. Gamboa, H. Zong, Y. Chen-Tsai, and L. Luo. Site-specific integrase-mediated transgenesis in mice via pronuclear injection. *Proc Natl Acad Sci USA*, 108(19):7902–07, 2011.
- J. Thomson and D. Ow. Site-specific recombination systems for the genetic manipulation of eukaryotic genomes. *Genesis*, 44(10):465–76, 2006.
- J. Thomson, R. Chan, R. Thilmony, Y. Yau, and D. Ow. PhiC31 recombination system demonstrates heritable germinal transmission of site-specific excision from the Arabidopsis genome. *BMC Biotechnol*, doi: 10.1186/1472-6750-10-17, 2010.
- H. Thorpe and M. Smith. In vitro site-specific integration of bacteriophage DNA catalyzed by a recombinase of the resolvase/invertase family. *Proc Natl Acad Sci USA*, 95(10):5505–10, 1998.
- H. Thorpe, S. Wilson, and M. Smith. Control of directionality in the site-specific recombination system of the Streptomyces phage phiC31. *Mol Microbiol*, 38(2):232–41, 2000.
- B. Thyagarajan, E. Olivares, R. Hollis, D. Ginsburg, and M. Calos. Site-specific genomic integration in mammalian cells mediated by phage phiC31 integrase. *Mol Cell Biol*, 21(12):3926–34, 2001.
- M. Tigges, T. Marquez-Lago, J. Stelling, and M. Fussenegger. A tunable synthetic mammalian oscillator. *Nature*, 457(7227):309–12, 2009.
- M. Tindall, E. Gaffney, P. Maini, and J. Armitage. Theoretical insights into bacterial chemotaxis. *Wiley Interdiscip Rev Syst Biol Med*, 4(3):247–59, 2012.
- G van Duyne. A structural view of cre-loxp site-specific recombination. *Annu Rev Biophys Biomol Struct*, 30:87–104, 2001.
- G van Duyne and K. Rutherford. Large serine recombinase domain structure and attachment site binding. *Crit Rev Biochem Mol Biol*, 48(5):476–91, 2013.
- E. Voit. *A first course in systems biology*. Garland Science, 2013.

- B. Wang, M. Barahona, and M. Buck. Engineering modular and tunable genetic amplifiers for scaling transcriptional signals in cascaded gene networks. *Nucleic Acids Res*, 42(14):9484–92, 2014.
- M. Watve, R. Tickoo, M. Jog, and B. Bhole. How many antibiotics are produced by the genus *Streptomyces*? *Arch Microbiol*, 176(5):386–90, 2001.
- B. Weinberg, N. Pham, L. Caraballo, T. Lozanoski, A. Engel, S. Bhatia, and W. Wong. Large-scale design of robust genetic circuits with multiple inputs and outputs for mammalian cells. *Nat. Biotech*, 35(5):453–462, 2017.
- J. Willey and A. Gaskell. Morphogenetic signaling molecules of the streptomycetes. *Chem Rev*, 111(1):174–87, 2011.
- M. Win and C. Smolke. Higher-order cellular information processing with synthetic RNA devices. *Science*, 322(5900):456–60, 2008.
- M. Woods and C. Barnes. Mechanistic modelling and bayesian inference elucidates the variable dynamics of double-strand break repair. *PLoS Comput Biol*, 12(10):e1005131. doi: 10.1371/journal.pcbi.1005131, 2016.
- M. Wu, R. Su, X. Li, T. Ellis, Y. Li, and X. Wang. Engineering of regulated stochastic cell fate determination. *Proc Natl Acad Sci USA*, 110(26):10610–15, 2013.
- Z. Xie, S. Liu, L. Bleris, and Y. Benenson. Logic integration of mRNA signals by an RNAi-based molecular computer. *Nucleic Acids Res*, 38(8):2692–701, 2010.
- Z. Xie, L. Wroblewska, L. Prochazka, R. Weiss, and Y. Benenson. Multi-input RNAi-based logic circuit for identification of specific cancer cells. *Science*, 333(6047):1307–11, 2011.
- Z. Xu, L. Thomas, B. Davies, R. Chalmers, M. Smith, and W. Brown. Accuracy and efficiency define Bxb1 integrase as the best of fifteen candidate serine recombinases for the integration of DNA into the human genome. *BMC Biotechnol*, doi: 10.1186/1472-6750-13-87, 2013.
- L. Yang, A. Nielsen, J. Fernandez-Rodriguez, C. McClune, M. Laub, T. Lu, and C. Voigt. Permanent genetic memory with >1-byte capacity. *Nat Methods*, 11(12):1261–66, 2014.

- L. Ye, J. Chang, C. Lin, Z. Qi, J. Yu, and Y. Kan. Generation of induced pluripotent stem cells using site-specific integration with phage integrase. *Proc Natl Acad Sci USA*, 107(45):19467–72, 2010.
- P. Yeh, M. Hegreness, A. Aiden, and R. Kishony. Drug interactions and the evolution of antibiotic resistance. *Nat Rev Microbiol*, 7(6):460–66, 2009.
- N. Yonemura, T. Tamura, K. Uchino, I. Kobayashi, K. Tatematsu, T. Iizuka, H. Sezutsu, M. Muthulakshmi, J. Nagaraju, and T. Kusakabe. PhiC31 integrase-mediated cassette exchange in silkworm embryos. *Mol Genet Genomics*, 287(9):731–39, 2012.
- N. Yonemura, T. Tamura, K. Uchino, I. Kobayashi, K. Tatematsu, T. Iizuka, T. Tsubota, H. Sezutsu, M. Muthulakshmi, J. Nagaraju, and T. Kusakabe. phiC31-integrase-mediated, site-specific integration of transgenes in the silkworm, *Bombyx mori* (Lepidoptera: Bombycidae). *Appl Entomol Zool*, 48(3):265–73, 2013.
- L. You, R. Cox, R. Weiss, and F. Arnold. Programmed population control by cell-cell communication and regulated killing. *Nature*, 428(6985):868–71, 2004.
- Y. Yu, Y. Wang, Q. Tong, X. Liu, F. Su, F. Quan, Z. Guo, and Y. Zhang. A site-specific recombinase-based method to produce antibiotic selectable marker free transgenic cattle. *PLoS One*, 8(5) e62457. doi: 10.1371/journal.pone.0062457, 2013.
- P. Yuan, K. Gupta, and G van Duyne. Tetrameric structure of a serine integrase catalytic domain. *Structure*, 16(8):1275–86, 2008.
- B. Zhang, L. Zhang, R. Dai, M. Yu, G. Zhao, and X. Ding. An efficient procedure for marker-free mutagenesis of *S. coelicolor* by site-specific recombination for secondary metabolite overproduction. *PLoS One*, 8(2) e55906. doi: 10.1371/journal.pone.0055906, 2013.
- F. Zhang, J. Carothers, and J. Keasling. Design of a dynamic sensor-regulator system for production of chemicals and fuels derived from fatty acids. *Nat Biotechnol*, 30(4):354–59, 2012.
- L. Zhang, X. Ou, G. Zhao, and X. Ding. Highly efficient *in vitro* site-specific recombination system based on streptomyces phage phiBT1 integrase. *J Bacteriol*, 190(19):6392–97, 2008.

- L. Zhang, L. Wang, J. Wang, X. Ou, G. Zhao, and X. Ding. DNA cleavage is independent of synapsis during *Streptomyces* phage phiBT1 integrase-mediated site-specific recombination. *J Mol Cell Biol*, 2(5):264–75, 2010.
- L. Zhang, G. Zhao, and X. Ding. Tandem assembly of the epothilone biosynthetic gene cluster by in vitro site-specific recombination. *Sci Rep*, doi: 10.1038/srep00141, 2011.
- Y. Zheng, J. Wu, Z. Chen, and M. Goodman. Chemical regulation of epigenetic modifications: opportunities for new cancer therapy. *Med Res Rev*, 28(5):645–87, 2008.
- X. Zou, L. Wang, Z. Li, J. Luo, Y. Wang, Z. Deng, S. Du, and S. Chen. Genome engineering and modification toward synthetic biology for the production of antibiotics. *Med Res Rev*, doi: 10.1002/med.21439, 2017.